4-15-2016 12:00 AM

Data Smoothing Techniques: Historical and Modern

Lori L. Murray
The University of Western Ontario

Supervisor
David Bellhouse
The University of Western Ontario Joint Supervisor
Duncan Murdoch
The University of Western Ontario

Graduate Program in Statistics and Actuarial Sciences
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Abstract

Some elementary data smoothing techniques emerged during the eighteenth century. At that time, smoothing techniques consisted of simple interpolation of the data and eventually evolved into more complex modern methods. Some of the significant milestones of smoothing or graduation of population data will be described including the smoothing methods of W.F. Sheppard in the early twentieth century. Sheppard's statistical interests focused on data smoothing, the construction of mathematical tables and education. Throughout his career, Sheppard consulted Karl Pearson for advice pertaining to his statistical research. An examination of his correspondence to Pearson will be presented and his smoothing methods will be described and compared to modern methods such as local polynomial regression and Bayesian smoothing models.

In the second part of the thesis, the development of Bayesian smoothing will be presented and a simulation-based Bayesian model will be implemented using historical data. The object of the Bayesian model is to predict the probability of life using grouped mortality data. A Metropolis-Hastings MCMC application will be employed and the results will then be compared to the original eighteenth-century analysis.

Keywords: Data smoothing methods, smoothing, splines, Bayesian smoothing, Sheppard, Pearson, life tables, history of statistics.

Acknowledgements

I would like to express my sincere thanks to my advisors Dr. David Bellhouse for his extensive knowledge in the history of probability and statistics and Dr. Duncan Murdoch for his expertise in statistical computation. I am genuinely grateful to have had the opportunity to study and research both the historical and modern methods of statistics with them.

I would also like to thank my husband Jeff and our three children for their love and support while working on this thesis.

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Introduction

In this thesis, historical and modern data smoothing techniques are presented and compared. The topics would be of interest to the statistical community and for those with an interest in the history of statistics.

Chapter 2 describes some of the significant milestones of early smoothing methods beginning in the late seventeenth century. Some of the earliest evidence of smoothing data can be found in the construction of life tables. Since population data contains irregularities, some adjusting or smoothing of the data is necessary. The collection of early data was often grouped by age and the gaps in the ages, such as deaths grouped into ranges with breaks at ages 10, 20, 30 years and so on, required smoothing for the purpose of interpolation. Elementary smoothing techniques such as visual interpolation, averaging, and mathematical interpolation were used to smooth out such irregularities in the data.

As the quality of data improved, smoothing methods became more advanced. Parametric and nonparametric models were developed along with graphical methods and difference formulas. The smoothing method of W.F. Sheppard in the early twentieth century was a significant milestone in the development of data smoothing. Sheppard's statistical career and correspondence to Karl Pearson are described in

Chapter 3. The correspondence spans three decades and it is obvious they became very close colleagues and good friends. Although the correspondence is one-sided (only the letters from Sheppard to Pearson are extant), they provide an interesting background to their statistical ideas and opinions before their manuscripts were published.

In the letters, Sheppard often asks Pearson for his advice regarding formulas for the tabulations of his tables related to the normal distribution. These were the first set of modern tables for the normal distribution based solely on the standard deviation. Throughout his career, Sheppard increased the accuracy of the tables by obtaining a higher number of decimals. Chapter 4 describes the methods of construction of his tables and how they were used.

Sheppard presented his smoothing method in a series of publications from 1912 to 1915. His method involves central differences and summation formulas based on least squares and is given in Chapter 5. We compare his method to modern smoothing methods such as local polynomial regression and Bayesian smoothing models.

The development of Bayesian smoothing and applications to the construction of life tables are given in Chapter 6.

In Chapter 7, a Bayesian smoothing model is developed to predict the probability of life using eighteenth-century mortality data. The model implements a Metropolis Hastings MCMC algorithm and the results are compared to the original eighteenth century analysis.

Chapter 8 provides a conclusion to the thesis. The various smoothing techniques presented in the thesis are summarized.

Chapter 2

The Development of Early Smoothing Techniques

2.1 Introduction

This chapter provides an overview of the development of early smoothing techniques beginning in the seventeenth century. Some of the earliest evidence of smoothing is found in the construction of life tables. A life table shows the number of persons alive at each age, and allows inferences to be made, such as the probability of surviving any particular age or the remaining life expectancy for persons at different ages. Population data contain irregularities and some adjusting or smoothing of the data is necessary in order to obtain reasonable estimates. The collection of detailed population data was slow to evolve. With the absence of a population census, early life tables were constructed from a limited number of observations spanning a short period of time. The compilers of early life tables did not disclose their exact methods of construction. However, given the techniques that were available to them at the time and examining others who used their methods, possible methods of construction will be described.

2.1.1 Graunt's Life Table

John Graunt, a London merchant, constructed a life table based on the observations recorded in the Bills of Mortality for the City of London, England. Starting in the early seventeenth century, the Bills of Mortality were bulletins published weekly to show the number deaths to warn residents of possible outbreaks of the bubonic plague. The London Bills consisted of the number of baptisms and deaths collected from parish clerks. As the main concern was for risk of recurrent epidemic diseases, only the cause of death was recorded and not the age at which a person died. Information about the collection and publication of these data can be found in "London Plague Statistics in 1665" (Bellhouse 1998). Using the London Bills, Graunt estimated the number of births and the number of persons living up to age 6, 16, and for every ten years up to age 86. He determined that for every 100 births, 36 die before the age of 6. Since the data was not grouped by age, Graunt had to guess the ages at which people had died given the cause of death. The results were published in 1662 in "National and political observations made upon the Bills of Mortality" (Graunt 1662) and are shown in Figure 2.1.

Table 2.1: Graunt's Life Table.

Age	*Number Alive*	*Deaths*
0	100	36
6	64	24
16	40	15
26	25	9
36	16	6
46	10	4
56	6	3
66	3	2
76	1	1
86	0	...

We observe a smooth progression after age 6 where the number of persons living is approximately equal to five-eighths of the previous one. This gives an annual survival

rate of about 95.4%, independent of age. The annual mortality rate according to Graunt's estimates would then be 1/18. The overall annual mortality rate shown in his data is 1/27 (Lewin and Valois 2003). Perhaps if Graunt had realized the discrepancy he would have adjusted the adult mortality rates to increase with age making the estimates in his table more accurate.

2.1.2 Halley's Life Table

Nearly thirty years later in 1693, Edmond Halley designed a life table based on mortality data for the valuation of life annuities. Casper Neumann, a Protestant pastor, collected the data from the parish registers in Breslau from 1687 to 1691. The city of Breslau in Silesia is now called Wroclaw in Poland. The data consist of the number of births and the number of deaths including the age at which people had died. Halley obtained the Breslau data, analysed it and constructed a life table. The Breslau data show that the population was approximately stationary. A stationary population is when the number of births equal the number of deaths and the age-specific mortality rates remain constant over time.

Analysing the data, Halley determined that the total population of Breslau was approximately 34,000 with a mean of 1238 births per year and 348 deaths in the first year of life. This gives (1238 + (1238-348))/2 = 1064, the mean number of infants alive in the first year. Halley rounded this number to start his population table with 1000 persons alive in the first year of age. Bellhouse (2011) illustrates how the additional 64 lives were redistributed throughout the early years of life. Halley's table is referred to as a life table, although by correct definition, it is a population table since it displays the mean number of persons alive at each age for Breslau (Greenwood 1941).

Table 2.2 shows Halley's estimates of the number of persons living at each age current from 1 to 84. Age current means a person is within that year of life but has not

reached their birthday of that year. Figure 2.1 show Halley's estimates of the number of persons alive as a function of age. Smoothing was necessary due to the irregularities of the data and the small numbers of deaths at the older ages. Calculating the slope for the yearly rates we find that Halley used piecewise linear interpolation to smooth out such irregularities. In general, Halley's estimates are approximately linear from age 12 to 78. The curve in Figure 2.1 is exactly linear between the dots.

Table 2.2: Halley's Breslau table.

Age. Curt.	*Per-sons.*	*Age. Curt.*	*Per-sons.*	*Age. Curt.*	*Per-sons.*	*Age. Curt.*	*Per-sons.*	*Age. Curt.*	*Per-sons.*	*Age. Curt.*	*Per-sons.*
1	1000	8	680	15	628	22	586	29	539	36	481
2	855	9	670	16	622	23	579	30	531	37	472
3	798	10	661	17	616	24	573	31	523	38	463
4	760	11	653	18	610	25	567	32	515	39	454
5	732	12	646	19	604	26	560	33	507	40	445
6	710	13	640	20	598	27	553	34	499	41	436
7	692	14	634	21	592	28	546	35	490	42	427
Age. Curt.	*Per-sons.*	*Age. Curt.*	*Per-sons.*	*Age. Curt.*	*Per-sons.*	*Age. Curt.*	*Per-sons.*	*Age. Curt.*	*Per-sons.*	*Age. Curt.*	*Per-sons.*
43	417	50	346	57	272	64	202	71	131	78	58
44	407	51	335	58	262	65	192	72	120	79	49
45	397	52	324	59	252	66	182	73	109	80	41
46	387	53	313	60	242	67	172	74	98	81	34
47	377	54	302	61	232	68	162	75	88	82	28
48	367	55	292	62	222	69	152	76	78	83	23
49	357	56	282	63	212	70	142	77	68	84	20

The results were published in 1693 in *Philosophical Transactions* titled "An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the City of Breslaw; with an attempt to ascertain the price of annuities upon lives" (Halley 1693). The table proved to be reliable in the valuation of annuities. Halley's table is considered the world's first life table and has been analysed extensively from a historical perspective since its publication (Bellhouse 2011a).

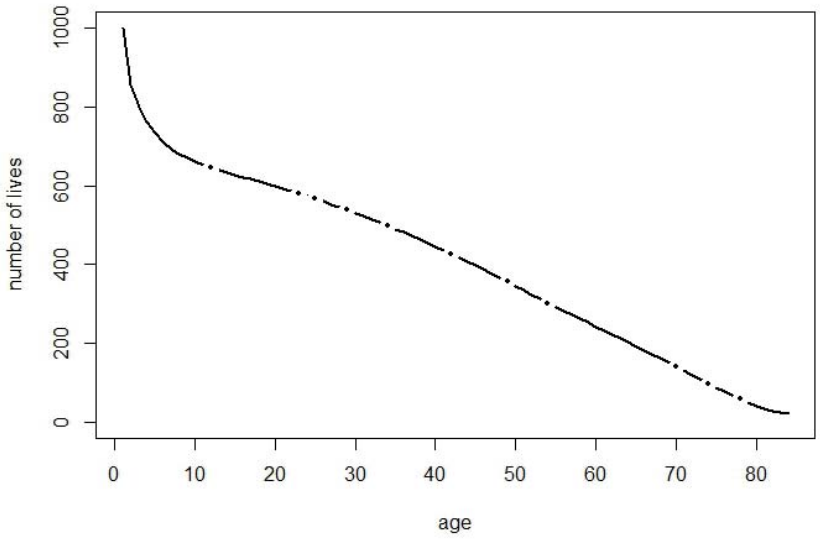

Figure 2.1: Halley's estimates of the number of lives at each age.

In the eighteenth century Halley's table was used as a source for other works such as Daniel Bernoulli's model on smallpox. Bernoulli used Halley's table in his calculations to model the probability of dying of smallpox. Adjusting the number of births from 1238 to 1300 was the only change Bernoulli made to Halley's estimates (Bacaër 2011, pp. 21–29).

2.2 Eighteenth-Century Smoothing

2.2.1 De Moivre's Survival Function

Deriving estimates for annuities using Halley's life table was a laborious task. In 1725 Abraham De Moivre developed a survival function that simplified the calculations. De Moivre used the fact that Halley's table was approximately linear after age 30 and used this assumption in deriving his survival probability model (De Moivre 1725). This allowed him to derive formulas for annuities of single lives. To approximate

annuities of joint lives as a function of the corresponding annuities on single lives, De Moivre used an exponentially decreasing function. The two assumptions, linear and exponential, are incompatible but De Moivre did it anyway to obtain a simple approximation (Bellhouse 2011b, pp. 161–164).

2.2.2 Smart's Life Table

Nearly 65 years after the publication of Graunt's life table, the collection of mortality data in London remained unchanged; parish clerks were required to report only the cause of death and not the age at which a person had died. In 1726 John Smart, a clerk at Guildhall London, wanted to construct a life table to estimate annuities but his design required the number of deaths at each age with observations taken over several years. Smart describes the problem in his book titled "Tables of Interest, Discount, Annuities, &c" (1726, p. 113). Smart didn't feel a change would be made during his lifetime. However, in less than two years after the publication of his book, parish clerks were required to include the approximate age of death. By 1737 Smart felt he had enough observations to construct a life table for the City of London.

Smart's life table along with the raw data is recorded on a broadside and held in the Guildhall Library in London (Smart 1738b). Extracted from the London Bills, the data gives the number of deaths for each year between 1728 to 1737 inclusive for each age group ranging from birth to greater than 90. Table 2.3 shows the total number of deaths for each age group and the corresponding number out of 1000.

Smart took the total number of deaths over ten years for each age group and determined the proportion out of 1000. We find from Smart's life table in Appendix A that the yearly rates remain constant over a few years. Smart retained the proportion of death for each age group given in the data and used piecewise linear interpolation to obtain the number of lives and deaths for the years between each age group. Figure 2.2 shows Smart's estimates for London (lower curve) and Halley's estimates

Table 2.3: Mortality rates from the Bills of Mortality for London, 1728 to 1737.

Age Group	*Total Deaths*	*Out of 1000*
0 to 2	103159	386
2 to 5	23505	88
5 to 10	9775	36
10 to 20	8242	31
20 to 30	19776	74
30 to 40	24302	91
40 to 50	23989	90
50 to 60	19693	74
60 to 70	16309	61
70 to 80	10684	40
80 to 90	6450	24
> 90	1266	5

for Breslau (upper curve). Smart estimates higher mortality rates than Halley's except for the older ages. Calculating the slope for the yearly rates we find Smart's curve is approximately linear from age 21 to 71. From age 21 to 60 the curves are nearly parallel. The curves are exactly linear between the dots.

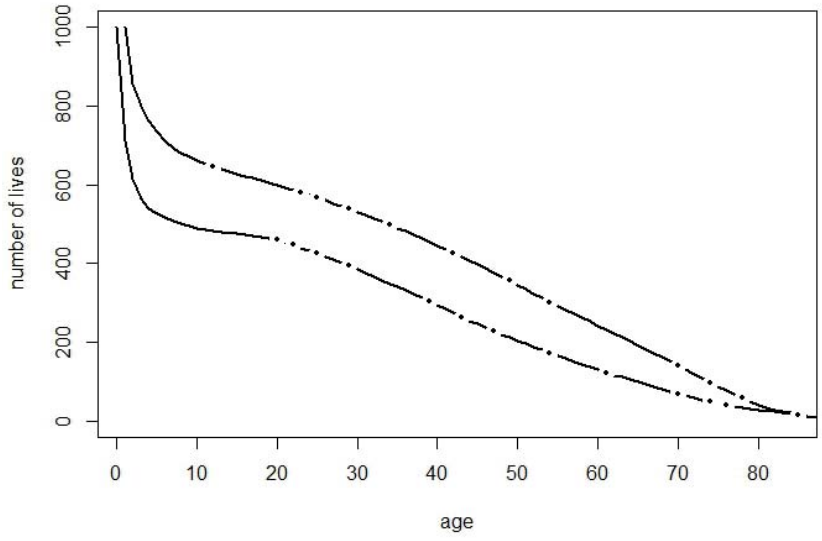

Figure 2.2: Smart's estimates (lower curve) and Halley's estimates of the number of lives at each age.

Smart wrote a letter to George Heathcote and enclosed a copy of his table estimates (Smart 1738a). Heathcote was a politician and a member of parliament in London. Dated February 25th, 1737, Smart explains to Heathcote how his life table is different than Halley's. Smart writes

> ...you will find a very great Difference more especially in the early part of Life. For 1238 Persons dying yearly at Breslau, the Doctor computes 616 of them, which is near one half, attain the age of seventeen: whereas by my Table, of 1000 Persons, there are but 501 who live to eight years of Age. But with respect to old Age, the Tables agree well enough for, by the one, 20 of the 1238, live to eighty four; by the other, 20 in 1000, to eighty three years of Age. (Smart 1738a)

Smart also explains how the two cities are different:

> ...Breslau is an inland City in Germany, inhabited chiefly by sober, industrious Peoples, Strangers to Luxury that Parent of all Vices, whereas London is a City abounding with Luxury amongst the Rich, and Debauchery amongst both of the Rich and Poor. (Smart 1738a)

Smart acknowledges that Breslau and London are different, but like Halley's table he assumed the population of London was stationary when he constructed the table. A consequence of assuming a stationary population is that the characteristics of the population are independent of time. This means that for each age group the number of live persons is always the same as that of the original life table. This is not realistic since most populations vary over time. A life table constructed with this assumption does not guarantee accurate estimates in the long run. The assumption was not practical for the city of London as it was with Breslau. At the time, London was experiencing significant immigration. Smart's estimates were based on the number of births, the number of deaths, and the age of death. He did not know the number of people in the population at each age, which made the table estimates unreliable.

2.2.3 Simpson's Life Table

The consequence of Smart's assumption of a stationary population was quickly recognized by Thomas Simpson. Simpson (1742) published a revision to Smart's table that tried to take into account migration. Simpson changed Smart's estimates up to age 25 and kept the remaining estimates the same. Simpson increased the number of births from 1000 to 1280, and using Halley's life table as a reference, used linear interpolation for the younger ages (Hald 1990, pp. 518–519). Figure 2.3 shows Simpson's estimates (lower curve) and Halley's estimates (upper curve) for the number of lives at each age. Simpson estimates higher mortality rates than Halley except for the older ages. Calculating the slope for the yearly rates we find Simpson's curve is approximately linear after age 12. The curves are nearly parallel from age 12 to 60. The curves are exactly linear between dots. Simpson's table was published in 1742 and used for insurance purposes (Hald 1990, p. 519).

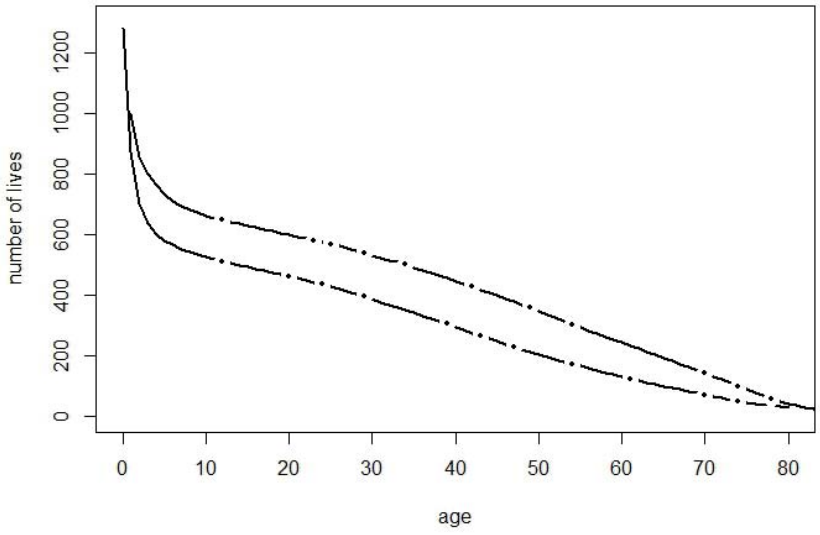

Figure 2.3: Simpson's estimates (lower curve) and Halley's estimates of the number of lives at each age.

2.2.4 The Northampton Table

Mathematician, philosopher and theologian Richard Price constructed a life table based on observations from the Register of Mortality at Northampton. The data are from the burial register of the Parish of All Saints in Northampton, England and spans 46 years from 1735 to 1780. The first version of the table was compiled using data from 1735 to 1770. With ten additional years of data in hand, Price revised and published the table in 1783 in *Observations reversionary payments; schemes for providing annuities for widows, and for persons in old age; on the method of calculating the values of assurances on lives.* The data consist of 4689 deaths and 4220 baptisms, a difference of 469 (or 10%). Price describes his method of construction on page 358 of his *Observations* although he is not explicit. William Farr (1848) explains in the 8th Report of the Registrar General that Price accounted for immigration at age 20. Based on Farr's description, W. Sutton (1883) proposes a method for the construction of the table. The construction of Price's life table is shown in Table 2.4.

Table 2.4: The Construction of the Northampton Table.

(1) *Age*	(2) *Deaths*	(3) *Deaths Adjusted*	(4) *Living*	(5) *Living 10000*	(6) *Less 1300 under 20*	(7) *Living Adjusted*	(8) *Northampton Table*
0	1529	1529	4689	10000	8700	11649.2	11650
2	362	362	3160	6739	5439	7283	7283
5	201	201	2798	5967	4667	6249	6249
10	189	189	2597	5538	4238	5675	5675
20	373	**351**	2408	5135	3835	5135	5132
30	329	**351**	2057	4387	...	4387	4385
40	365	365	1706	3638	...	3638	3635
50	384	384	1341	2860	...	2860	2857
60	378	378	957	2041	...	2041	2038
70	358	358	579	1235	...	1235	1232
80	199	199	221	471	...	471	469
90	22	22	22	47	...	47	46
100	...	...	0	21	...	0	0

The first two columns in Table 2.4 show the data with the number of deaths for

each age group. For example, there are 1529 deaths from birth up to age 2, and there are 362 deaths from age 2 up to age 5, and so on. Price smooths the number of deaths by averaging the age groups 20 to 30 and 30 to 40 so that they are equal (shown in bold). Column 4 corresponds to the number of person living if the population was stationary with the initial value for the number of persons alive from birth to age 2 being the sum of all the deaths from column 3. Column 5 is the number of persons living for each age group proportionally increased for a population size of 10,000. Column 6 is smoothed to account for immigration by decreasing up to age group 20 to 30 in column 5 by 1300 (13% of 10,000 instead of the 10% suggested by the data). Column 7 increases the first five age groups by the proportion (5135/3835) required to restore the age group 20 to 30 to the original value in column 5. The last column shows Price's Northampton table. The differences between the last two columns differ by no more than 3 between the age 20 and 90.

Price's Northampton table was constructed properly based on the given data (Registrar General 1848, p. 291). However, Farr (1853) states that the data did not accurately represent Northampton because there were a great number of Baptists living in the town and they do not baptize infants. This reduced the ratio of christenings to deaths, which decreased the average life expectancy. The consequence of this was that the mean duration of life was assumed to be 24 years when it was really about 30 years. The table was used by the Equitable Life Assurance Society and the British government for 20 years to determine the price of annuities it sold. This led to losses since the longevity of the annuitants was greater than what the table indicated.

2.3 Nineteenth-Century Smoothing

2.3.1 The Carlisle Table

Joshua Milne employed a graphical smoothing method in the construction of the Carlisle Table, a life table based on data from the City of Carlisle. Milne was an actuary for the Sun Life Assurance Society. The table was published in 1815 in *A Treatise on the Valuation on Annuitities and Assurances on Lives and Survivorships* (Milne 1815). The data were provided by John Heysham, a medical doctor, and was taken from population data and the Bills of Mortality of two parishes in Carlisle. The data consist of a census of grouped data for the number of persons living for the years 1780 and 1787. The data include the number of deaths for the same age groups with birth to 5 given in one year intervals covering the period from 1779 to 1787.

Columns 1 to 3 in Table 2.5 show the data with the number of persons alive for each age group for the years 1780 and 1787 respectively. The total number of persons living for the eight-year period is calculated as the sum of the 1780 and 1787 censuses multipled by 4 and is shown in column 4. Column 5 is the total number of persons living (column 4) divided by the width of each age group and rounded to the nearest integer.

Milne begins his graphical approach by constructing rectangles whose base corresponds to the widths of each age group and the heights as calculated in column 5 of Table 2.5. For example, the age group birth to 5 has 8772 persons living over a five year period which represents the area of the first rectangle with the height given by 8772/5=1754 from column 5. Using his knowledge and experience, Milne drew a smooth continuous curve through the tops of the rectangles such that any additional area added to the rectangles was equal to the amount removed. Milne knew to start the curve high because the infant mortality data showed a high number of deaths in the first year of life.

Table 2.5: The Carlisle data.

(1) *Age Group*	(2) *Population in 1780*	(3) *Population in 1787*	(4) *Living 8 year total*	(5) *8 year total at each age*
0 to 5	1029	1164	8772	1754
5 to 10	908	1026	7736	1547
10 to 15	715	808	6092	1218
15 to 20	675	763	5752	1150
20 to 30	1328	1501	11316	1132
30 to 40	877	991	7472	747
40 to 50	858	970	7312	731
50 to 60	588	665	5012	501
60 to 70	438	494	3728	373
70 to 80	191	216	1628	163
80 to 90	58	66	496	50
90 to 100	10	11	84	8
100 to 105	2	2	16	...

Figure 2.4 is Milne's graph of the Carlisle population curve (Milne 1815, p. 101). Milne used the graph for the purpose of illustration but did not include the values for the horizontal and vertical axes.

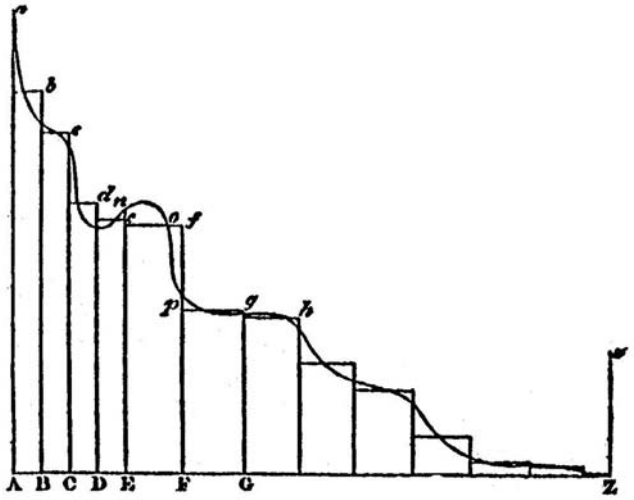

Figure 2.4: Carlisle population curve.

The Carlisle table is constructed from the graph in Figure 2.4. The number of persons living for each year is determined by finding the year on the horizontal axis and the corresponding value on the curve. The same graphical interpolation method can be used to find the number of deaths as illustrated by actuary George King (1883).

The Carlisle table was widely adopted by actuaries and used for many years for the valuation of annuities (BMJ 1902). The *British Medical Journal* featured Milne's method in its 1902 publication concluding that the graphical method "is simpler, more elegant, and equally accurate with the analytical method" (BMJ 1902).

2.3.2 Gompertz-Makeham Law of Mortality

Mathematician and actuary Benjamin Gompertz derived a parametric model for the construction of life tables. The idea of the model was first introduced and published in *Philosophical Transactions* in 1820 and 1825, and further developed and presented to the Royal Society in 1861 (Gompertz 1820, 1825, 1861). The model is known as the Law of Mortality. Let D_x be the cumulative number of deaths up to age x, then

$$D_x = Bc^x \tag{2.1}$$

where B and c are constants. Fellow actuary William Makeham revised the model to improve the accuracy. The model was published in the *Journal of the Institute of Actuaries* in 1859 titled "On the Further Development of Gompertz's Law". Makeham's model includes the addition of a constant term A and is given by

$$D_x = A + Bc^x. \tag{2.2}$$

The model is useful for smoothing mortality observations and for calculating the value of life insurance (Hald 1990, p. 513).

2.3.3 The English Table

Farr constructed the first four English Life Tables. The third life table was published in 1864 and the method for its construction is described in full in his book, *English Life Table* (Farr 1864). The data are based on the 1841 and 1851 population censuses for England and Wales and the number of deaths for the 17 years from 1838 to 1854 for both males and females from the civil registrations. The data consist of population and deaths for individual years from birth to age 4, for every five years up to age 15, and for every ten years up to greater than 95 (Farr 1864, p. xix).

Farr obtained a uniform distribution of deaths using

$$p_x = \left(\frac{2 - m_x}{2 + m_x}\right) \tag{2.3}$$

where p_x is the probability that someone age x will survive to age $x + 1$ and m_x is the number of persons dying at age x divided by the mean population at age x. In other words, m_x is the rate at which people are dying in the middle of the year of age x to $x + 1$ and is formally known as the central force of mortality. Farr retained the rates of mortality for ages under 5 given in the data. For the 10-year groups Equation (2.3) gives the force of mortality for integral ages instead of for the mean of the year of age.

Farr assumed that the force of mortality (instantaneous mortality rate at age x) for a country increased in a geometrical progression using the relation $\mu_{x+t} = r^t \mu_x$ for t years and $r^{10} = \mu_x + 10/\mu_x$. Then

$$-\ln(p_x) = \int_0^1 \mu_{t+1} dt = \frac{r - 1}{\ln(r)} \mu_x. \tag{2.4}$$

Transforming into common logarithms we have

$$-\log(p_x) = \frac{k^2(r - 1)}{\log(r)} \mu_x. \tag{2.5}$$

where $k = \log_{10}e$. The values for $\log(p_x)$ for ages 3, 4, 7, 12 and every ten years thereafter were used as the basis for third difference interpolation after dividing the table into sections. Table divisions were done separately for males and females based on the analysis of the data. Farr (1864, p. clxvii) obtained the deaths rates for each year of life and tabulated the results. The yearly mortality rates are given in logarithms for both male and female from birth to age 109. The computations involved were extensive and the tables were used for insurance purposes.

2.4 Early Twentieth-Century Smoothing

A special edition on data smoothing methods was published in 1921, *Tracts for Computers* by E.C. Rhodes and edited by Karl Pearson. This rare publication examines and compares some of the data smoothing (or graduation) techniques in use at the time and served in part as the motivation for this thesis. A large amount of experimental and observational data were collected during W.W.I, which prompted serious discussion on smoothing methods (Rhodes 1921, p. 4). The staff of the Galton Laboratory, UCL were engaged in research for the Admiralty Air Department and Ministry of Munitions and the collection of wartime data was related to fuses, elasticity, propellers, aircraft and ballistics trajectories, and range tables (Galton Laboratory Wartime Research Papers, UCL Special Collections, Pearson/9). The smoothing methods of John Spencer (1904), W.S.B. Woolhouse (1869), A. Cauchy (1837), T. Sprague (1886) and W.F. Sheppard (1914b) are considered. Spencer's graduation formula, also known as the summation formula, uses 15 or 21 values tabulated in order to obtain one smoothed value at a time. The process is repeated with the series of smoothed values proceeding by constant third differences. The method is simple for practical use and was widely used by actuaries. Woolhouse's method uses 15 values to smooth out the central value, using repeated summations until each value is smoothed. He was the first person to use differences to smooth data. His method assumes that the third differences in the given series are negligible and uses parabolic

interpolation. Cauchy suggests a method of smoothing observations using a known function, and Sprague uses a graphical approach using osculatory interpolation which requires previous knowledge and experience of the given data.

Great attention is given to Sheppard's smoothing method using differences based on least squares. His method is proven to perform well by having the smallest or same magnitude of mean square error as the other methods studied in the tract. Sheppard's statistical career and correspondence to Pearson is given in Chapters 3, the construction of his tables in Chapter 4, and his smoothing method in Chapter 5.

2.5 Conclusion

The collection of detailed population data increased and was recorded over longer periods of time. Elementary smoothing techniques evolved into more complex modern methods. The progression of smoothing methods began with visual interpolation, averaging, and mathematical interpolation, and developed into smoothing methods using parametric and nonparametric models, differences and graphical methods. The motivation for developing new methods or improving on existing ones is to find a way to adjust the data that results with smoothed values that are closer to the true values, and thus reducing the error, while keeping in mind that the new method is suitable for practical use.

Advanced smoothing methods are employed in the construction of modern life tables. For example, the construction of life tables in use by Statistics Canada (2015) involve two methodologies: logistic models and splines. B-splines are used for smoothing the ages of death due to their flexibility. The logistic model replaced the quadratic model in 2005. Studies show that the mortality rate in countries with higher quality data tended to follow a logistic curve (Statistics Canada 2015). The process of smoothing population data continues to be refined as the quality of data improves.

Chapter 3

The Correspondence from Sheppard to Pearson

3.1 Background

William Fleetwood Sheppard was born in 1863 in Sydney, Australia. He attended grammar school in Brisbane and was sent to England to finish his education at Charterhouse School. He won a scholarship to Trinity College, Cambridge and was Senior Wrangler in the Mathematical Tripos of 1884. Sheppard became a Fellow of Trinity College and published a paper on Bessel functions (Sheppard 1889). Sheppard left Cambridge to pursue a career in law but returned to his interest in education and research, and focused on statistics. In 1896, he was appointed Junior Examiner in the Education Department and later promoted to Assistant Secretary. He retired in 1921 at the age of 58. He then became a Senior Examiner at the Univeristy of London before moving to Edinburgh in 1926. He worked at the Edinburgh University and was elected a Fellow of the Royal Society of Edinburgh in 1932 (Sheppard 1938).

3.2 Early Correspondence

At the beginning of his statistical career, Sheppard consulted British statistician Karl Pearson, a leading pioneer of modern statistics who could provide Sheppard with statistical advice and expertise. Sheppard wrote a letter to Pearson describing a manuscript he was working on with Francis Galton and asked if he would review it when it was completed. This was the first letter of a series of 23 letters that are archived at University College, London (Pearson 1896–1926). The letters have been transcribed and can be found in Appendix B. For reference, a brief description of the letters are given in Table 3.1.

Table 3.1: Descriptions of the 23 letters from Sheppard to Pearson.

Letter No.	*Date*	*Description*
1	3 June 1896	Sheppard describes an unfinished paper he has been working on with Galton.
2	16 June 1896	Sheppard asks Pearson of any possible employment opportunities.
3	19 October 1896	Sheppard questions Pearson regarding his method for finding the moments of a polygon.
4	20 October 1896	Sheppard questions Pearson on his methods on the fitting of curves.
5	10 May 1899	Short discussion on Sheppard's work on a quadrature formula. Sheppard offers Pearson a mathematical problem for student examinations.
6	31 March 1900	Sheppard asks Pearson if he has any use for his quadrature formula. He tells Pearson that his paper on normal correlation will be published.
7	5 April 1900	More discussion with regards to Pearson using Sheppard's quadrature formulae and the fitting of curves.
8	8 April 1900	Quadrature formulae for volumes and for specific curve-types.
9	4 May 1900	The fitting of curves and the organization of a future paper that Pearson is working on.
10	7 May 1900	Brief discussion on the fitting of the "cloudiness curve."

11	18 December 1900	Sheppard tells Pearson he would like to write a short article about interpolation formulae for surfaces.
12	13 Febrary 1901	Sheppard encourages Pearson to write an article on the mathematical treatment of statistics for the *Times*.
13	16 February 1902	Discussion as to where Sheppard's tables related to the normal distribution could be published. Personal writing about Pearson's health.
14	13 February 1908	Sheppard asks Pearson the proper etiquette for reusing questions from others' examination papers in textbooks.
15	18 May 1911	Sheppard discusses a problem on probability.
16	4 October 1911	Sheppard's tables related to the normal distribution including a table of values.
17	23 July 1915	Personal topics regarding Sheppard's family and the health of his eldest son.
18	18 October 1916	Brief discussion about a probability problem.
19	10 April 1925	Brief discussion about Sheppard's tables related to the normal distribution with some calculated values.
20	6 September 1925	Discussion about Sheppard's tables related to the normal distribution and extending the number of decimal places.
21	26 November 1925	More discussion about Sheppard's tables with some calculated values.
22	2 December 1925	Sheppard desribes 3 of his tables and includes them in the letter.
23	29 June 1926	Details on Sheppard's tables related to the normal distribution. A personal note on Pearson's operation.

The letters span three decades from 3 June 1896 to 29 June 1926. The majority of the correspondence spans the first decade during which time many of Sheppard's papers were published. The letters begin very formally with "Dear Sir" and discussions about statistical methods, but over the course of the thirty years they become informal where Sheppard speaks of his family and of Pearson's health. It is obvious that they became very good colleagues and good friends.

In general, the letters pertain to specific papers that Sheppard was working on

with the hope of being published. He shares his statistical ideas and methods with Pearson and frequently asks for his advice on which journal would be the most suitable for the publication of his manuscripts. Reducing the costs of publishing his methods and tables was also considered. For example, Sheppard suggests using the derivatives of a function instead of the differences for his tables to save time and space. Sheppard estimated his quadrature formula would take up 10 or 11 pages using octavo-sized paper. The speed of the calculations are discussed throughout the correspondence. Sheppard knew how many minutes it would take to use his method of interpolation using a Brunsviga mechanical calculator that he had on loan from the Royal Society. Sheppard wondered if Charles Vernon Boys, a British physicist and inventor, could devise a machine to simplify the process of calculating large numerical determinants. Vernon Boys (1944) designed and constructed an integration machine. Instruments, such as a planimeter were used for testing what Sheppard called "closeness of fit." A planimeter determines the area of a two-dimensional shape.

Sheppard was also interested in teaching. He asks Pearson of any possible employment opportunities and the average hourly rate for private mathematical coaching. In the early correspondence, Pearson offered Sheppard a position to teach astronomy. However, Sheppard declined stating it wasn't the best subject for him when he attended Cambridge. Sheppard knew that Pearson set examinations and offered a problem that he could use in his examinations. He asks Pearson if questions from publications are allowed to be used in student examinations. The correspondence shows the solutions to probability questions that Sheppard had worked out for future examinations.

The letters suggest that Pearson had a major influence on Sheppard's statistical work. Sheppard compares his methods and results to Pearson's in a non-competitative way to try to fully understand the statistical concepts. For example, Sheppard discovered that Pearson had published a paper on the normal correlation based on the multiple integral (the multivariate normal distribution). He had worked out a method

for normal correlation for the double integral (the bivariate normal distribution) and found that his method was different than Pearson's. Sheppard did not always agree with Pearson's methods and would offer an explanation as to why. Sometimes he would give an alternate method and ask for Pearson's opinion. Sheppard shared proofs and formulas and referenced Pearson in his published works. Pearson must have liked Sheppard's results. He referenced Sheppard's methods and formulas such as his corrections of moment estimates for normally grouped data and his quadrature formulas in his own published works (Pearson 1902, 1914a, 1914b). Details of Pearson's references will be given later in this chapter.

The letters provide a rare and insightful glimpse into the personal and professional relationship between Sheppard and Pearson. They give some of the background details of Sheppard's methods and formulas that would eventually be published and adopted by other statisticians.

3.3 Statistical Correspondence

The main theme of their correspondence was the fitting of curves but they also discussed probable error formulas, moment estimates and corrections to moment estimates for grouped normal data, quadrature formulas, tests of fit, Pearson's chi-square test, and tables for the normal density function.

Before proceeding to the specific topics discussed in the correspondence, it is important to describe some of the statistical terminology and theory that was being developed at the time. Towards the end of the nineteenth century, asymmetrical distributions were becoming accepted and new distributions were being developed to model skewed data. Previously, it was assumed that all continuous statistical data were normally distributed. Probability distribution functions were called frequency distributions or curves of frequency. In 1895, Pearson developed four types of frequency curves to model skewed observations (Pearson 1895). By 1916, the number of

curves had increased to twelve and they became well known as the "Pearson Family of Frequency Curves" (Stigler 2008). The details of how Pearson derived his family of curves can be found in §2.3.3. Although they were referred to as parameters, the constants of the frequency curves were not parameters in the way we define them today. Quantities such as the mean and standard deviation were expressed, when possible, in terms of the frequency constants.

Pearson sometimes used the constants of his frequency curves as though they were parameters but this proved to be consequential. Historian Stephen Stigler (2008) explains why. Referring to an 1898 paper jointly authored by Pearson and colleague L.N.G. Filon (Pearson and Filon 1898), Stigler describes a major error when they incorrectly derived an asymptotically approximate multivariate normal distribution for the errors of estimation from the expansion of a log-likelihood ratio. The source of the error was the substitution of integrals for sums in the Taylor expansion. Stigler points out this was equivalent to replacing the sums with expectations. The Taylor expansion they used was about the estimates meaning the expectations were then functions of the estimates and not that of the true values. In other words, there was no distinction between the estimates and the parameters of the model. Stigler explains the consequence of the error:

> All the expectations are computed as if the estimated values were true values, and the result is a distribution for errors that does not in any way depend upon the method used to estimate. (Stigler 2008)

Unfortunately, Pearson lacked the notion of a distribution of true values of the parameters and "for him there was no 'true value,' only a summary estimate in terms of observed values" (Stigler 2008). At the time, the consequences of the method went unnoticed. The idea of parametric modelling was not introduced until 1922 by R.A. Fisher. Fisher presented a method for fitting curves using maximum likelihood estimation. The new method proved to be superior to Pearson's method since the maximum likelihood estimators are asymptotically unbiased consistent, efficient and

asymptotically normal.

3.3.1 Probable Error

The first letter in the Pearson Papers collection describes an unfinished paper on the normal curve that Sheppard and Galton had been working on. Sheppard explains how the paper is entirely theoretical and geometrical without the use of any differentiation or integration. The paper contains new material on the correlation between normal distributions and that non-normal distributions would only be considered for the purpose of analysing them into component normal distributions. Sheppard writes that the paper takes up a great deal of space but he wanted to treat the subject thoroughly. Sheppard wanted to know if Pearson would be willing to look at the paper when it was finished and if he might suggest a suitable journal for its publication. Sheppard wondered if the paper had a chance of being published in the *Philosophical Transactions*. Given the date of the letter and the subject of the unfinished paper, it appears Sheppard was referring to his paper, "On the geometrical treatment of the 'normal curve' of statistics," dated October 1897 and published in the *Proceedings of the Royal Society of London* (Sheppard 1897b). The paper was revised and republished under the title, "On the application of the theory of error to cases of normal distributions and normal correlations," in 1899 in *Philosophical Transactions of the Royal Society* (Sheppard 1899c). In the paper, Sheppard makes reference to Galton, highlighting his contribution on normal correlation. Sheppard includes his proof of a theorem in bivariate normal correlation, which is now sometimes known as Sheppard's theorem on median dichotomy (MacKenzie 1981, p. 97). In addition, the paper includes methods for evaluating probable error for the frequency constants of the normal distribution and tables for calculating probable error.

The term "probable error" was first used in the early nineteenth century to describe what we now call the median error of an estimate (Stigler 2008). If m is the probable error and σ is the standard deviation, then the probable error is

$m = 0.6745\sigma$. The first and third quantiles of a normal distribution are 0.6745σ from the mean. The probability that a deviation is greater than the probable error is 0.5 and is equal to the probability of a deviation less than the probable error. If the observed deviation is less than 3 times the standard error it is approximately equivalent to the observed deviation being less than 4.5 times the probable error.

In his 1899c paper, Sheppard gives two applications where probable error can be used: for computing the discrepancy between the observed values and the true values, and for hypothesis tests. The hypothesis tests include the test for normality, test for normal correlation and the test for independence of two distributions. Generally speaking, for about half the values of X, the discrepancy, d, should be less than the probable discrepancy, q, and amongst the remaining values the discrepancy should not be a large multiple of the probable discrepancy. The ratios, d/q, are computed to determine if they are or are not greater than we might reasonably expect. Sheppard includes a table of values to compare with the computed ratio values. The method is similar to the rejection region approach for hypothesis tests that is used today. Let q be the quartile deviation (probable discrepancy) and m the number of random values. If the area of the standard normal distribution between the points $x = -p/q$ and $x = +p/q$ is ϕ, then the probability of at least one of the values of δ being greater than p is $1-\phi$. If ϕ is chosen such that the probability is 0.5, the corresponding value p will be the "probable limit" of δ. The tables gives 20 values for m corresponding to the values of the ratio p/q. For example, when m=1 then p/q=1 and when m=10, p/q=2.716. For values greater than m=20, Sheppard (1899, p. 123) suggests using Chauvenet's criterion for the rejection of one out of $m/\ln(4+1/2)$ observations. William Chauvenet was an American mathematician and astronomer.

Sheppard gives several examples to illustrate the hypothesis tests using probable error. For example, a hypothesis test to determine if a distribution is normal is given using grouped data of the chest measurements of 5,732 local Scottish militia, a famous dataset from the *Edinburgh Medical and Surgical Journal* (1817, pp. 260-263). The

first step is to calculate the mean $\bar{x}$, and standard deviation s. Sheppard (1897a) uses a special formula to calculate the standard deviation based on grouped data that he derived in a previous paper. He uses areas to derive the variance which he calls the mean square and is similar to the shortcut formula we use today to calculate the variance for grouped data $[\sum(fx^2) - (\sum(fx))^2/n)]/(n-1)$ where f is the group frequency. This was Sheppard's first published paper in statistics where he developed corrections to moment estimates for normally grouped data. Details about the paper can be found in §3.2.2.

In the next step, Sheppard creates new bins of the chest measurements to equal the midpoint of each class, for example, 33 belongs to the bin 32.5 to 33.5 and 48 belongs to the bin 47.5 to 48.5. He then computes the class-index, α_i, for each value which represents the standardized proportion for each class $[(2n_i/n) - 1]$ where $n_i = \sum_{j=1}^{i} f_j$. The middle ten values (35.5 to 44.5) are standardized, z_k for $k = 1, \ldots 10$. Sheppard then calculates $\bar{x} + sz_k$ for each class k. The discrepancy values, d_k, are the differences between each midpoint value and $\bar{x}+sz_k$. Let $\phi(z_k)$ be the standard normal pdf evaluated for each class k, then the standard deviation for each discrepancy is $[s^2(1-\alpha_k^2)/4\phi(z_k)^2 - (1+\frac{1}{2}z_k^2)]^{1/2}/\sqrt{n}$, which when multiplied by 0.67449 gives the probable discrepancy values q_k. For the ten classes, four of the actual discrepancies are less than the probable discrepancies, and the remaining six are greater. In addition, the ratios $(d/q)_k$ are compared to Sheppard's probable limit δ, mentioned above, for m=10. Nine of the ten values are less than the corresponding p/q, and therefore, it is concluded that the data appear to justify the hypothesis of a normal distribution.

The probable error can be calculated using Table V on pages 159 to 166 of Sheppard's 1899c paper. The tables contain values for the mean square (variance), denoted by N, and the intermediate values (shown in the table between two values of N) that correspond to the probable values, $\mathrm{Q}\sqrt{\mathrm{N}}$, where Q=0.67448975... For example, if the variance is calculated as N= 0.019300, then the value of $\mathrm{Q}\sqrt{\mathrm{N}}$ to three decimal places is 0.094.

At the time, probable error was used as a measure of the variability of the constants of frequency curves resulting from a random sample. In this case, the probable error is the standard deviation of the constant multiplied by 0.67449. The convention of using probable error as a measure of goodness of the sample, rather than the standard deviation, was adopted since the theory was developed from the normal curve. At the end of the first letter, Sheppard wrote a post script stating he "should be much gratified if any of my work would be of use to you in your own investigations." In their 1898 paper, Pearson and Filon derived the probable error for the frequency constants but used a different method than Sheppard (Pearson and Filon 1898). They used a Bayesian approach with a uniform prior, which was sometimes referred to as the Gaussian method (Stigler 2008, p. 5). It was a method of inverse probability and was commonly used over the nineteenth century. As noted by Stigler, Pearson and Filon's derivations contained some errors in the distinction between the estimates and the population parameters. Over time, Pearson distanced himself from his probable error methods in preference for Sheppard's non-Bayesian methods (MacKenzie 1981, pp. 203-204) and (Stigler 2008). Pearson referenced Sheppard as being the "fundamental memoirs on the subject" in the editorial appearing in the volume of his 1903 paper titled *On the probable error of frequency constants* (Pearson 1903, p. 35). Pearson included Sheppard's methods for finding the probable error of the frequency constants for five types of his system of curves. Sheppard's methods for evaluating the probable error for the frequency constants involve simple linear functions of frequency counts using a Taylor expansion when necessary. The probable errors are estimated and then the moments are found from the variances and covariances of the counts.

3.3.2 Corrections of Moment Estimates

The early correspondence includes references to methods for finding moment estimates. In Letter 3, Sheppard enclosed a manuscript for Pearson to review, suggesting it might be suitable for the Royal Statistical Society. Sheppard states that he will

put the mathematical part into a separate paper and asks for Pearson's advice on the possibility of it being published in the *Philosophical Magazine* or the *Cambridge Philosophical Society*. It appears Sheppard is referring to his first statistical paper published in 1897 in the *Journal of the Royal Statistical Society* (Sheppard 1897a). The paper summarizes his corrections of moment estimates for normally grouped data. They were fully presented mathematically in 1898 (Sheppard 1898) and became known as "Sheppard's corrections" (Aitken 1938).

For continuous frequency distributions, it can be assumed frequencies are centered at the midpoints of the class intervals when calculating the moments. This introduces some error and corrections are required. In modern notation, let μ_n be the nth central moment, μ_n^* the corresponding corrected moment and c the bin width. Sheppard's first five corrected moments are:

$$\begin{aligned}
\mu_1^* &= \mu_1 = 0 \\
\mu_2^* &= \mu_2 - \frac{1}{12}c^2 \\
\mu_3^* &= \mu_3 \\
\mu_4^* &= \mu_4 - \frac{1}{2}\mu_2 c^2 + \frac{7}{240}c^4 \\
\mu_5^* &= \mu_5 - \frac{5}{6}\mu_3 c^2
\end{aligned}$$

In a memoir, after Sheppard's death, A.C. Aitken highlights an error where Pearson incorrectly omits the use of the corrections in his 1895 paper (Pearson 1895). Aitken writes that corrections of the moment estimates should be applied in a certain case for grouped data and gives Sheppard credit for deriving them (Aitken 1938). Aitken describes how Sheppard was tactful in pointing out Pearson's error and because of this, his corrections were not universally adopted for some time.

Pearson used Sheppard's corrections of moment estimates throughout his 1902 paper, "On the Systematic Fitting of Curves to Observations and Measurements" (Pearson 1902). In 1914, Sheppard's corrections were used in an illustration in the

Tables for Statisticians and Biometricians, a publication edited by Pearson (Pearson 1914a, 1914b). The first four moments are calculated on the head circumferences of 1,306 criminals. The data consist of 40 sub-groups and Pearson suggests that 20 sub-groups be used instead, and that Sheppard's corrections would fully adjust for the difference (Pearson 1914a, p. lxxvi).

3.3.3 Methods of Fitting Curves

Differences in their statistical views and methods began to surface a few months into their correspondence. Two back-to-back letters (Letters 3 and 4) reveal some of these differences. In Letter 3, Sheppard informs Pearson that after reading his essay on "Skew Variation", he modified a manuscript he was working on to include a reference to Pearson's paper and for illustration to include one of his tables. Sheppard (1898) was working on his manuscript on corrections of moment estimates which includes an appendix on the moments of a polygon based on a frequency curve. In the letter, Sheppard writes that his method for finding the moments of a polygon based on observations is very different from Pearson's and that his method "seems the more correct." This would have been of interest to Pearson since the moments of a polygon based on observations were used in his method for fitting frequency curves to data. Sheppard is referring to Pearson's 1885 paper titled, "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material" (Pearson 1895). Pearson's essay on "Skew Variation" was his first paper where he gives a systematic method for the theoretical fitting of curves. As mentioned earlier, it was at this time that asymmetrical distributions were becoming accepted for fitting statistical data and Pearson was a leader in the development. His system of frequency curves was derived using the following method. A density function, $f(x)$, is defined as a solution to the differential equation:

$$\frac{df}{dx} = \frac{(x-a)f(x)}{b_0 + b_1 x + b_2 x^2}. \tag{3.1}$$

The differential equation is based on the logarithm of the density function of the normal distribution and the probability mass function of the hypergeometric distribution. The sign of the roots of the characteristic equation in the denominator determine two main types of curves each containing sub-type curves. The types of curves relate to the values of the parameters. To find the values of the parameters, Pearson used the method of moments. He imported the method from physics (mechanics) (Porter 2004, p. 240). In mechanics, a "moment" is a measure of force about a point of rotation (center of mass) and is the product of the magnitude of the force by its perpendicular distance from the point. In statistics, the first four moments represent the mean, dispersion of measurements around the mean, skewness and kurtosis. The parameters in the denominator of the differential equation are expressed in terms of the moments of the frequency curves. The values of the parameters determine the curve type. Any Pearson curve can be uniquely determined by the first four non-central moments if they exist. The nth non-central moment is

$$\mu'_n = \int_{-\infty}^{\infty} x^n f(x) dx \tag{3.2}$$

and the nth central moment about the mean μ of the distribution is

$$\mu_n = \int_{-\infty}^{\infty} (x - \mu)^n f(x) dx \tag{3.3}$$

Using a standard conversion formula,

$$\mu_n = \sum_{j=0}^{n} \binom{n}{j} (-1)^{n-j} \mu'_j \mu^{n-j} \tag{3.4}$$

the non-central moments can be converted to central moments. Pearson derived the parameters in the denominator of the differential equation in terms of the central

moments:

$$\begin{aligned} b_0 &= -\sigma^2(4\beta_2 - 3\beta_1)/D, \\ a = b_1 &= \beta_1^{1/2}\sigma(\beta_2 + 3)/D, \\ b_2 &= (2\beta_2 - 3\beta_1 - 6)/D \end{aligned} \tag{3.5}$$

where $\beta_1 = \mu_3^2/\mu_2^3$, $\beta_2 = \mu_4/\mu_2^2$, $\mu_2 = \sigma^2$ and $D = 10\beta_2 - 12\beta_1 - 18$. The moments of the frequency curves are approximated by a formula derived by Pearson. He begins by constructing rectangles based on observations shown in Figure 3.1. The figure is taken from page 346 of his 1895 paper (Pearson 1895).

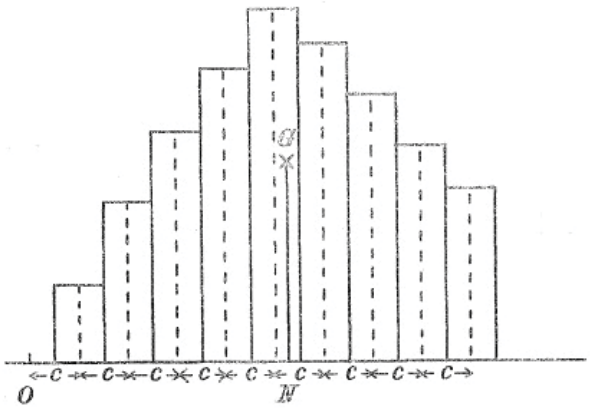

Figure 3.1: Pearson's histogram.

Pearson defines y_r as the height of the rth rectangle and c as the distance between the midpoints of each rectangle. Polygons are formed by joining the tops of the midpoints of adjacent rectangles to form a frequency curve shown in Figure 3.2. Pearson refers to the frequency curve as the "curve of observations." The diagram of the curve is from page 349 of his 1895 paper. The ordinates y_1, y_2, y_3, ... , $y_r, y_{r+1}, \ldots$, are the frequencies of deviations falling within the ranges $x_1 \pm 1/2c$, $x_2 \pm 1/2c$, $x_3 \pm 1/2c$, $\ldots$, $x_r \pm 1/2c$, and so on. The area of the polygon is approximately equal to that of the curve, and the first non-central moments of the two areas are also approximately equal. A Taylor expansion is used to approximate the non-central moments. Pearson's

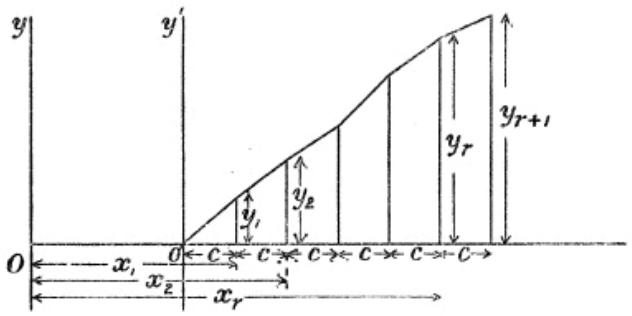

Figure 3.2: Pearson's diagram of a frequency curve based on observations forming a series of polygons.

nth non-central moment of a series of polygons based on observations is

$$\mathrm{M}'_n = \sum_{r=1}^{p} \left[2y_r \left(\frac{x_r^n c}{2!} + \frac{n(n-1)}{4!} x_r^{n-2} c^3 + \frac{n(n-1)(n-2)(n-3)}{6!} x_r^{n-4} c^5 + \dots \right) \right]. \quad (3.6)$$

Pearson derives the first five sample non-central moments from Equation 3.6 and then converts them to sample central moments. The sample central moments are used to estimate the true moments via Equation 3.5. The values of the parameters in Equation 3.5 are approximated using the data and used to determine the type of Pearson curve.

In Letter 3, Sheppard explains to Pearson that his methods for finding the moments of a polygon based on observations give different results and offers an explanation as to why his method seems more correct. To avoid confusion, I will use η to define the number of observations, since Sheppard and Pearson used n to represent both the number of observations and the nth moment. Pearson takes the ordinate y_r as proportional to η_r. Sheppard states that this works for a first order approximation using rectangles, but a polygon is equivalent to a second order approximation. Taking $y_r = \eta_r$, it underestimates where the curve is concave to the base and overestimates where the curve is convex to the base, resulting in a larger standard deviation. In other words, when constructing polygons a correction must be made. To make it a second order approximation the ordinate should be $\eta_r + \frac{1}{2}[\eta_r - \frac{1}{2}(\eta_{r-1} + \eta_{r+1})]$

rather than η_r. Sheppard tells Pearson that he will reference his method in his paper but "will give the corresponding formula for the nth moment accurately (your M'_n)." Sheppard is referring to Pearson's formula (3.6).

Sheppard's viewpoint becomes clear in his derivation of the moments of a polygon in the appendix titled, "Moments of a Polygon" of his 1898 paper (Sheppard 1898, pp. 378–380). In the paper, Sheppard restates what he wrote in the letter highlighting the issue as to why a larger standard deviation occurs when the ordinate is proportional to number of observations. The polygon lies inside the spurious (empirical) curve when the spurious curve is convex, and outside it where it is concave. Therefore, the mean square error of the polygon is greater than the empirical curve and the mean square error of the empirical curve is greater than the true curve. Sheppard references Pearson's derivation of Equation 3.6 in his paper but includes a correction to estimate the curve more accurately (Sheppard 1898, p. 379). He also gives an alternate formula to derive the moments based on areas using the formula, $A_r = \frac{1}{2}h(z_{r-1} + z_r)$, rather than the ordinates.

The conversation continues in a second letter written the next day (Letter 4). It appears that Sheppard is writing in response to an earlier letter from Pearson. Sheppard thanks Pearson for providing an explanation for his method on curve fitting and writes that he had misunderstood it. Pearson was fitting a curve that was not a frequency curve but was related to a frequency curve. Despite the clarification, Sheppard was still perplexed. Referring to the same diagram (Figure 3.2) mentioned in the letter the previous day, Sheppard tries to explain the issue. He acknowledges that the ordinate y_r of the curve at every point x_r is proportional to the area of the curve between $x \pm \frac{1}{2}c$, but states "there is no finality in this." Sheppard's concern is that when this is applied to a normal distribution, for example, the shape of the new curve will depend on the value of c. This in turn affects the degree of smoothing and thus, the fitted curve. Sheppard continues to explain how he looks at the fitting of curves differently by offering an alternate view:

> Apparently we look at the thing in different ways. I do not try to find a frequency curve which the numbers given could be successive areas: I try to find the frequency curve which would result if the causes or whatever they are–which regulate the particular magnitude in the individuals measured acted in the same way on an infinite number of individuals, the hypothesis being that the particular individuals are a chance selection from this infinite number. (Pearson 1896–1926)

Sheppard is suggesting that his method finds the theoretical frequency curve rather than the empirical frequency curve. In his 1899 paper, Sheppard geometrically derives the normal curve and comments in a footnote how his method is different than Pearson's method in reference to Figure 3.2. Sheppard acknowledges that Pearson's curve of observations converges to the normal curve when n is made "indefinitely great" (Sheppard 1899c, pp. 120–122).

Subsequently, Sheppard derived a quadrature formulae to approximate the population moments in Equation 3.2 as an alternate method to Pearson's sample moments in Equation 3.6. We know from the correspondence that Sheppard spent a great amount of time deriving the quadrature formulae. In 1902, Pearson references Sheppard's quadrature formulae in a paper on the fitting of curves to observational data (Pearson 1902). More details on the correspondence and the development of Sheppard's quadrature formulae are given in the next subsection (§2.3.4).

3.3.4 Quadrature Formulae

In response to Pearson, Sheppard accepts his request to work on the development of quadrature formulae. Sheppard informs Pearson that he had tried to write out a formula and to apply it to his curves but found it wasn't as good as calculating the area using a direct method (Letter 5). However, several months later, Sheppard appears to have worked out the derivations of the quadrature formulae and given them to Pearson

(Letter 6). His quadrature formulae include an extension to volumes that he thought would be useful for others such as naval architects. Naval architects calculate volumes of complex shapes (ships' hulls) to determine displacement. He informs Pearson that he would like to have them published. Sheppard briefly describes the sketch of a paper and asks Pearson if he would like to incorporate his quadrature formulae into one of his own papers as well (Letter 7). A few days later, Sheppard suggests that the two of them meet in person to discuss the quadrature formulae (Letter 8). The quadrature formulae were published a few months later (Sheppard 1900), and in 1902, Pearson referenced them in his own paper and included examples (Pearson 1902). Pearson refers to the formulae as "Sheppard's Rule" and compares it to other quadrature rules available at the time such as Simpson's Rule. Pearson writes:

> Accordingly Mr Sheppard has determined the best coefficients for the corrections to the chordal and tangential areas when one, two or three differences only are used. He has provided the following quadrature formulae which seem to me of much interest and practical value. (Pearson 1902, page 275)

Pearson concludes that Sheppard's formula gives the best approximation when fitting frequency curves to statistical data. Sheppard's quadrature formulae provides a way to use Pearson's method of moments for the fitting of curves, which gives comparable results to the method of least squares. Pearson suggests that his method of moments be used when the method of least squares is too laborious or impractical (Pearson 1902, p. 271).

3.3.5 Tests of Fit and Pearson's Chi-Square Test

In Letter 7, Sheppard writes to Pearson about his method for estimating the accuracy of fit. Sheppard states that Pearson's method of using percentages rather than probable error seems unsatisfactory. He questions Pearson by stating "if 6% is good, for 500

observations, surely it may be bad for 2000?" He continues to ask what determines "what is good & what is not" since "what is good for a curve with 3 constants may be bad for a curve with 5 constants." As mentioned, nineteenth-century statisticians used the terms constants and parameters interchangeably. Sheppard was also concerned with the misfits at different points of the curve since they are not always of equal weight.

Sheppard sent a lengthy letter to Pearson describing his views on testing the fit of a curve (Letter 9). In Sheppard's words, "test of misfit" describes what we call today "goodness of fit". In the letter, Sheppard offers an example of fitting a curve with n classes or bins using an equation with $n-1$ constants. Sheppard explains that if a good fit is achieved you may believe that you have captured the underlying population distribution, but then if, say, the number of classes are doubled, it may no longer be a good fit and thus, may not have captured the underlying population distribution. Sheppard suggested that if the curve is fitted with less than $n-1$ constants, there will be many solutions. It appears that Sheppard is asking Pearson how to determine which fit is best. He continues to discuss life tables and suggests that depending on the application it may be desirable to have a good fit in certain parts of the curve, say the beginning rather than in other parts of the curve.

Sheppard then proposes an order of topics for a paper that appears to be in progress. The three main topics are (1) data manipulation which included smoothing and interpolation, (2) goodness of fit where Sheppard proposed his modification to Pearson's test of misfit known today as Pearson's chi-square test, and (3) analysis such as variation and the calculation of moments. Sheppard had a great deal of knowledge of these statistical topics. Goodness of fit problems are discussed in his 1899 paper. A few years later he published a paper in three parts on the development of smoothing methods for fitting curves (Sheppard 1914a, 1914, 1915). Sheppard's least squares smoothing method was published in 1921 in a book on smoothing methods titled, *Tracts for Computers*, which was edited by Pearson (Rhodes 1921). Sheppard later

published a paper where he derived Pearson's chi-square test (Sheppard 1929). Interestingly, Pearson's famous paper on the chi-square test (Pearson 1900) was published only a few months after the date of this letter.

3.3.6 Numerical Tables

Throughout the correspondence and more frequently towards the end, Sheppard discusses the details of his mathematical tables. In Letter 13, Sheppard suggests that some of the tables were useful for various purposes. The letter corresponds directly to his publication of tables in 1903 (Sheppard 1903). The first two tables give the values for $\Phi(x)$ for $x = 0.00$ to $x = 6.00$ using steps of 0.01 up to 7 and 10 decimal places. The third and fourth tables give the values for x and the corresponding values of the evaluated probability density function, $\phi(t)$, of the standard normal distribution in terms of α where $\alpha = 2\int_0^x \phi(t)dt$. Sheppard states how the first table is useful for small values of x and the second table is useful for large values of x, and therefore, could be used for various purposes such as calculating moments of an area. Sheppard explains why the tabulations for the area in the third table stop at $\alpha = 0.80$. If it went beyond this value, the differences would increase making the calculations unmanageable. He also suggests that his tables could be used for interpolation and determining probable error in testing for normality, but that Pearson might prefer his chi-square test instead. It appears that Sheppard decided to omit his table on probable errors that he mentioned in the letter and instead include a short discussion using an example in the introduction on how to calculate probable error.

The publication containing the four tables based on the standard normal curve became popular among statisticians. In 1907, Galton included the third table in his paper (Galton 1907). Pearson was also interested in Sheppard's tables (Pearson and Lee 1908, for example p. 61 & 65) and was the editor for their publication in the 1914 (Volumes 1 and 2) and 1924 editions of *Tables for Statisticians and Biometricians* (Pearson 1914a, 1914b, 1924). We know from the correspondance

from Fisher to Sheppard that statisticians at the Galton Laboratory had access to an unpublished version of Sheppard's tables since they needed a higher decimal accuracy when working on problems that required rigorous conclusions. Sheppard hoped to eventually have these tables extended and published as well. Additional details on Sheppard's tables and his methods of construction are given in Chapter 4.

3.4 Later Correspondence

As the years progressed, the letters became more personal. Sheppard discussed details about his family and shows concern for Pearson's health. The longevity of their relationship suggests that Pearson was highly influential in Sheppard's statistical career. The correspondence offers an interesting background into their personal views and opinions regarding their work before their papers were published. They provide insight into the development of their methods at a pivotal time in modern statistics.

Chapter 4

Sheppard's Tables

4.1 Background

Sheppard held a lifelong interest in the construction of tables related to the normal distribution. They were the first set of modern tables for the standard normal distribution based solely on the standard deviation, i.e. x/σ is used as the argument. Prior to this, they used a modulus, $x/(\sigma\sqrt{2})$, or probable error. Initially, the calculations were carried out to 5 decimal places at wide intervals and were published in 1899 (Sheppard 1899c). The tables were extended to 7 and 10 decimal places and published in 1903 (Sheppard 1903). The tables were widely used and reproduced unchanged in successive issues in *Tables for Statisticians and Biometricians* with Pearson as editor (Pearson 1914a, 1914b, 1924).

4.2 The Construction of Sheppard's Tables

Using Sheppard's notation, the probability density function is defined as

$$z_x = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{4.1}$$

and the upper tail area of the normal curve is

$$\frac{1}{2}(1-\alpha_x) = \frac{1}{\sqrt{2\pi}}\int_x^{\infty} e^{-\frac{t^2}{2}}dt \tag{4.2}$$

and the lower tail area of the normal curve is

$$\frac{1}{2}(1+\alpha_x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt \tag{4.3}$$

where

$$\alpha_x = 2\int_0^x z_t dt \tag{4.4}$$

which is known as the error function, erf(x), defined as the probability of a random normal variable with mean equal to 0 and variance equal to $\frac{1}{2}$ on the range $-x$ to x.

Sheppard (1903, pp. 180-181) outlines his methods of constructing four tables in his 1903 publication. The first two tables in the paper give the area (4.3) and the probability density function (4.1) to 7 decimals in terms of x at intervals of 0.01 starting at 0.00 to 4.50. The tables are extended to 10 decimals for values of x from 4.50 to 6.00. The first and second differences are also tabulated. Sheppard constructed the tables by quadrature and central difference formulas that he developed (Sheppard 1899b). For example, let $\phi(x)$ be the probability density function. For values of $x = 0$ up to 2.50 the quadrature formula (midpoint integration) is given by

$$\int_x^{x+h} \phi(t)dt = \left(1 + \frac{1}{24}\delta^2 - \frac{17}{5760}\delta^4 + \dots\right) h\,\phi\left(x + \frac{1}{2}h\right) \tag{4.5}$$

where $h = 0.01$ and for any function $f(x)$ the first and second central differences are

$$\delta f(x) = f(x + \frac{1}{2}h) - f(x - \frac{1}{2}h)$$

and

$$\delta^2 f(x) = f(x+h) - 2f(x) + f(x-h) \tag{4.6}$$

and so on. The probability function $\phi(x)$ is evaluated for the intermediate values, $x = 0.005$, 0.015, 0.025, and so on, by successive multiplication rather than interpolation. Every tenth value is checked using F.W. Newman's table (1883) of the function e^{-y} evaluated for values of y. The values for $x = 0.01$, 0.02, 0.03, and so on, are obtained and the corresponding values for $\phi(x)$ are found by interpolation. The remainder of the tables, for x greater than 2.50, the values of $\phi(x)$ are calculated using the function $\log_{10}\phi(x)$ and the integral is evaluated using the quadrature formula (Sheppard 1903, p. 180).

The third and fourth tables in the paper give the values of x to 7 decimals and the probability density function (4.1) to 7 decimals in terms of the area (4.4) at intervals of 0.01 from 0.00 to 0.80. As mentioned in Chapter 3, Sheppard ended the table at 0.80 because the calculations become unmanageable beyond this point.

Sheppard provides a list of uses for the four tables. The first and second central differences are given for the purpose of interpolation. Sheppard developed a method for extending the accuracy of the tabulations (1899a). His method shows how to interpolate between the values of x using differences and smoothing to obtain a higher number of decimals when required. An example is given in his 1903 publication to show the process. Inverse interpolation is also discussed using examples. Inverse interpolation is used when the area is known and the standardized variable is required. Sheppard suggests the tables can be used for tests for normality and for calculating correlation volumes. An example showing Sheppard's test for normality is given in Section §2.3.1. The product of the tabulated probability density function (4.1) for two independent variables (the bivariate normal distribution for two independent variables) can be used in calculating correlation volumes.

4.3 The Probability Integral

Sheppard continued to construct new tables for the standard normal curve until the end of his life. It was Sheppard's wish to construct a set of tables to have "as many decimal places as would ever be required" (Sheppard 1939). Several years were spent on improving the accuracy by using a higher number of decimals. He consulted Pearson for advice regarding his tabulations and asked for his recommendation as to who might be interested in publishing his set of tables having a high number of decimal places.

When Pearson retired in 1933, Ronald Fisher took over the Galton Laboratory. He would then have access to Sheppard's tables. Fisher was still clearing out some of Pearson's material in 1936 (Fisher Box 1978, p. 346). At that time, Fisher wrote to Sheppard on behalf of the British Association Mathematical Tables Committee for his permission to publish the tables for the normal distribution (Fisher 1936). Fisher knew that the seven-decimal place table was published but was interested in publishing a higher accuracy version such as the one that was available at the Galton Laboratory. In the letter, Fisher writes:

> It has been felt by a great many people to be a great pity that the full table was never published, for it would have been exceedingly useful on many occasions. (Fisher 1936)

Fisher indicated that the use of these tables would be valuable for the construction of other tables. Unfortunately, Sheppard died before the tables were ready for publication. Fisher continued to work on getting the tables published with Sheppard's son, N.F. Sheppard (Fisher 1937). They were eventually published in 1939 in a volume titled, *The Probability Integral*, prepared by the British Association on Mathematical Tables (Sheppard 1939).

The volume contains six tables related to the standard normal distribution in-

cluding differences and derivatives. A summary of the formulas for the six tables contained in the volume are shown in Table 4.1

Table 4.1: Sheppard's tables related to the normal curve.

Table I	$\frac{1}{2}(1-\alpha_x)/z_x$ for $x=0$ to 10 by h=0.01 to 12D
Table II	$\frac{1}{2}(1-\alpha_x)/z_x$ for $x=0$ to 10 by h=0.10 to 24D
Table III	$-ln(\frac{1}{2}(1-\alpha_x))$ for $x=0$ to 10 by h=1.00 to 24D
Table IV	$-ln(\frac{1}{2}(1-\alpha_x))$ for $x=0$ to 10 by h=0.10 to 16D
Table V	$\log_{10}(\frac{1}{2}(1-\alpha_x))$ for $x=0$ to 10 by h=0.10 to 12D
Table VI	$\log_{10}(\frac{1}{2}(1-\alpha_x))$ for $x=0$ to 10 by h=0.01 to 8D

where h is the step-size and D is the number of decimal places.

Table I is the ratio of the tail area of the normal curve to its bounding ordinate, with reduced derivatives, at intervals of one-hundredth of the standard deviation, to twelve decimal places.

Table II is the ratio of the tail area of the normal curve to its bounding ordinate, with reduced derivatives, at intervals of one-tenth of the standard deviation, to twenty-four decimal places. Since the calculations contained a large number of decimals, the table was constructed using Laplace's continued fraction (Sheppard 1939). Laplace's continued fraction is given by

$$e^{y^2}\int_y^\infty e^{-u^2}du = \frac{1}{2y}\Bigg/\left(1+\cfrac{1/2y^2}{1+\cfrac{2/2y^2}{1+\cfrac{3/2y^2}{1+\cdots+\cfrac{n/2y^2}{1+\cdots}}}}\right) \tag{4.7}$$

If we let $y = x/\sqrt{2}$ and $u = t/\sqrt{2}$ we have

$$e^{x^2/2}\int_x^\infty e^{-t^2/2}dt = \cfrac{1}{x+\cfrac{1}{x+\cfrac{2}{x+\cfrac{3}{x+\cdots+\cfrac{n}{x+\cdots}}}}} \tag{4.8}$$

The function (4.2) is used for the calculations in Table I and Table II. Stated in the introduction of the publication,

> Sheppard used the fact that any tabular entry is the sum of the next tabular entry and its reduced derivatives, all taken positively, while the reduced derivatives of any entry are simple linear functions, with known coefficents, of these same quantities. (Sheppard 1939)

Following Sheppard's method, Table I is constructed by subtabulating Table II to the interval 0.01. Sheppard's subtabulations were completed by Mr. F.H. Cleaver under the direction of the Mathematical Tables Committee of the British Association using the Association's National accounting machine. Cleaver was the association's first "computer" and was appointed in January, 1938 (Croarken and Campbell-Kelly 2000). Using the fundamental values from Table II, Sheppard constructed Tables III, IV, and V.

Table III is the negative natural logarithm of the tail area of the normal curve, for integral multiples of the standard deviation, to twenty-four decimal places and Table IV is the negative natural logarithm of the tail area of the normal curve, with reduced derivatives, at intervals of one-tenth of the standard deviation, to sixteen decimal places. The function

$$L(x) = -\log_e \frac{1}{2}(1-\alpha_x) \tag{4.9}$$

is used to obtain the values for Table III and Table IV.

Table V is the common logarithm of the tail area of the normal curve, with reduced derivatives, at intervals of one-tenth of the standard deviation, to twelve decimal places and Table VI is the common logarithm of the tail area of the normal curve, with second central differences, at intervals of one-hundredth of the standard deviation, to eight decimal places. The function

$$l(x) = \log_{10}\frac{1}{2}(1 - \alpha_x) \tag{4.10}$$

is used to obtain the values for Table V and Table VI.

Table IV gives the second central differences. Tables I, II, IV, and V include derivatives for interpolation. The nth reduced derivative of a function is defined as $f_n(x) = h^n f^{(n)}(x)/n!$. The number of reduced derivatives given varies from 3 to 16. Sheppard used h as the argument interval so that accurate interpolation could be obtained using a Taylor expansion.

The Mathematical Tables Committee felt that a table for eight decimal places would be useful and appointed statistician H.O. Hartley to calculate the tabulations for Table VI. Hartley followed Cleaver in June 1938 to become the association's second "computer." Hartley obtained a Ph.D. in mathematics in 1934. He studied at Humboldt-Universität zu Berlin before going to England to escape the Nazis. Two of his earliest papers were on computational methods. The advantage of using the logarithm function is to obtain a higher number of significant digits.

All of the tables were checked for accuracy under the direction of the British Association Committee. Table III was checked by direct calculation and Tables II, IV, and V by summing the function and its reduced derivatives for each value of the argument and comparing the result to the next value. It was noted in the publication that not one error was found in Sheppard's calculations confirming his remarkable precision and dedication. The volume includes the following statement given by the

Association Committee:

> The Committee, in issuing this volume, believe that the completion and publication of his tables of the probability integral constitute just that memorial to Sheppard's unsurpassed labours in the field of Mathematical Statistics, which he would himself most greatly have appreciated. (Sheppard 1939)

Sheppard's goal for publishing a set of tables having as many decimal places as would ever be required was finally accomplished.

4.4 How Sheppard's Tables Were Used

Sheppard's tables were widely used by statisticians and users of statistics for many years. They were used for a wide range of applications such as tests for normality, the fitting of curves, the construction of other tables such as probable error tables, and the calculation of multivariate normal distributions. As mentioned in Chapter 3, the tables were also used for calculating the moments of an area. Some examples on how to use the tables are provided by Pearson in *Tables for Statisticians and Biometricians* (Pearson 1914a). Abstracts of the tables have been published in books by statisticians and scientists. For example, a graphical approach is used in fitting a normal curve to data (Brown 1921, p. 43). Using the values extracted from one of Sheppard's tables, a normal curve was superimprosed onto a histogram of bisection data. Similarily, the tables were used to compare the areas between graduated and ungraduated curves using Endowment Assurances data (Elderton and Johnson 1969, pp. 72–73). The same approach is achieved today using statistical software.

William Gosset used Sheppard's tables when he introduced the t-distribution under the pseudonym "Student" published in *Biometrika* in 1908 (Student 1908, p. 24). Gosset's employer, Guinness Breweries, did not permit Gosset to publish his work under his own name. In the paper, Gosset compares the probability of the yield of

corn per acre between the t-distribution and the normal distribution. The fact that he referred to them as "Sheppard's tables" shows how commonplace the tables were at the time. This would be the same as a reference to a well known method such as a Taylor expansion in a modern paper.

We know from the correspondance from Fisher to Sheppard that statisticians at the Galton Laboratory had access to an unpublished version of the tables to a higher decimal accuracy (Fisher 1936). A higher number of decimals would give the Laboratory statisticians the increased level of accuracy when working on problems that required rigorous conclusions. The Laboratory was "most heartily" thankful for the twelve and sixteen figure tables on loan from Sheppard (Pearson 1914b, p. 10).

In a review of Sheppard's publication, Hartley indicates that a high degree of accuracy is required for a variety of scientific problems and in order that rigorous conclusions may be drawn, a high decimal accuracy of the normal curve is essential (Hartley 1940). For example, a generalization of Airy's theory of absorption spectra leads to the following integral

$$\int_0^{x_1} \cos(ax - b\,\alpha_x)dx. \tag{4.11}$$

The integral must be evaluated for large values of the constants a and b. Clearly, the high number of decimals in Sheppard's tables would be useful using Sheppard's values for α_x defined in Equation (4.4).

The tables continued to be published in successive issues of *Tables for Statisticians and Biometricians* by K. Pearson (Pearson 1914a, 1924) and *Biometrika Tables for Statisticians* from 1954 to 1970, by E.S. Pearson and H.O. Hartley (Pearson and Hartley 1970).

Chapter 5

Sheppard's Smoothing Methods

5.1 Background

In the early twentieth century, Sheppard (1912) developed a polynomial smoothing method. As mentioned in Chapter 2, some of the smoothing methods in use at the time were developed by Spencer (1904), Woolhouse (1869), Cauchy (1837), Sprague (1886). Woolhouse and Spencer's methods were based on local quadratic fitting. Sheppard's method was based on local polynomial fitting.

In 1912, Sheppard presented his smoothing method using central differences. Two years later, he presented an alternate computation method in a series of four papers using central summations (Sheppard 1914a, 1914b, 1914c, 1915). The alternate method gives the same result as the method using central differences but is less computationally intensive if the dataset is large. Sheppard's smoothing curve can also be obtained using successive application of the method of ordinary least squares. Sheppard methodically derives his methods with extensive detail and works through some practical applications for illustration. In this chapter, Sheppard's smoothing methods are described and compared to modern smoothing techniques.

5.2 Sheppard's Smoothing Formula in Terms of Central Differences

We begin with a set of predictor variables x corresponding to a set of response variables u. We consider a sequence of equally-spaced values of x corresponding to a sequence of values of u. Smoothing consists of replacing the sequence, $\ldots$, u_{-1}, u_0, u_1, $\ldots$ with another sequence $\ldots$, v_{-1}, v_0, v_1, $\ldots$ such that each v is a "linear compound" of the corresponding u and n others on each side of it. The smoothed value is

$$v_0 = p_n u_n + p_{n-1} u_{n-1} + \cdots + p_{-n} u_{-n}. \tag{5.1}$$

where p_n, p_{n-1}, $\ldots$, p_{-n} are coefficients. The expression for v_0 is symmetrical about u_0. Sheppard uses the term "linear compound" instead of the term "linear function" because the method is concerned with the proportion in which the various u's have to be compounded in order to produce the required result rather than with functionality. The problem is to find a polynomial v in x of degree j,

$$v(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_j x^j. \tag{5.2}$$

such that

$$\sum_{i=1}^{m} (u_i - v_i)^2 \tag{5.3}$$

is a minimum for a set of m observations, u_i is the set of unsmoothed values and v_i are the corresponding set of smoothed values. Sheppard developed a general solution to the problem in terms of central differences. The central differences of u_0 are calculated using Table 5.1.

Table 5.1: Central differences of u_0.

x	u	δu	$\delta^2 u$	$\delta^3 u$	$\delta^4 u$
$\vdots$	$\vdots$				
x_{-2}	u_{-2}	$\vdots$			
		$u_{-1}-u_{-2}$	$\vdots$		
x_{-1}	u_{-1}		$u_0-2u_{-1}+u_{-2}$	$\vdots$	
		u_0-u_{-1}		$u_1-3u_0+3u_{-1}-u_{-2}$	$\vdots$
x_0	u_0		$u_1-2u_0+u_{-1}$		$u_2-4u_1+6u_0-4u_{-1}+u_{-2}$
		u_1-u_0		$u_2-3u_1+3u_0-u_{-1}$	$\vdots$
x_1	u_1		$u_2-2u_1+u_0$	$\vdots$	
		u_2-u_1	$\vdots$		
x_2	u_2	$\vdots$			
$\vdots$	$\vdots$				

It is assumed that the errors of the sequences between the u's and v's are independent and have the same mean square error. Since the problem consists of $j+1$ equations of condition, the necessary condition for replacing u_0 by v_0 is that $u_0 - v_0$ should only involve differences of u_0 of orders of $j+1$ and upwards. This means that the smoothed or adjusted value, v_0, can be formed by adding these differences. The expression in Equation (5.1) is symmetrical about u_0, so it only involves u_0 and its central differences of even order. Equation (5.1) is rewritten as

$$v_0 = q_0 u_0 + q_2 \delta^2 u_0 + \cdots + q_{2n} \delta^{2n} u_0 \tag{5.4}$$

where the q_0, q_2, ..., q_{2n} are coefficients, and $\delta^2 u_0 \ldots \delta^{2n} u_0$ are the even central differences of u_0. The relation between the coefficients in Equation (5.1) and the coefficients q's in Equation (5.4) is obtained using binomial coefficients and the coefficients derived from them. The coefficients are chosen by using the mean square error for v_0

subject to the j+1 conditions, and the smoothing formula becomes

$$v_0 = u_0 + (-1)^k \frac{1 \cdot 3 \ldots (2k+1)}{1 \cdot 2 \ldots k} \sum_{s=k+1}^{s=n} \frac{(s-1)(s-2)\ldots(s-k)}{(2s+1)(2s+3)\ldots(2s+2k+1)} (n+\tfrac{1}{2}, 2s] \delta^{2s} u_0 \tag{5.5}$$

The notation for the brackets is calculated as $(n+\frac{1}{2}, 2s] = (n-s+1)(n-s+2)\ldots(n+s)/(2s)!$. The formula can be used for degree j=2k+1 or j=2k and gives a single smoothed value for each set of m=2n+1 u's. The degree of the polynomial can vary throughout the dataset, and Sheppard suggests grouping the given values together to determine which order should be used. The number of values on each side of u_0 is called the bandwidth and is denoted by n. Since there are a specified number of u's, and the degree, j, is known, the coefficients in (5.5) can be derived. Sheppard performed the laborious task of deriving all of the coefficients that he thought would ever be required for practical purposes. He constructed a set of tables of the coefficients and included them in his paper (Sheppard 1912, pp. 378–382). For example, if there are m=13 values of u, and the degree j=2, the coefficients for the smoothing formula are: 1, +0, −6, −8, $-\frac{45}{11}$, $-\frac{12}{13}$, $-\frac{1}{13}$.

The table of central differences of u_0 is extended and the process is repeated to find the next smoothed value. For instance, if m=13, the first smoothed value, v_7, is obtained using u_7 and its even central differences. The second smoothed value, v_8, is obtained using u_8 and its even central differences, and so on.

The smoothed values for the ends of the range, e.g. v_1, v_2, …, v_6, are obtained using the formulas for the central differences of v_0. For degree j=2k+1 or j=2k, the formula for the even central differences of v_0 is

$$\delta^{2t} v_0 = (-1)^{k-t} \frac{[t+\frac{1}{2}, k+1]}{t!(k-t)!(n+\frac{1}{2}, 2t]} \sum_{s=0}^{s=n} \frac{s(s-1)\ldots(s-k)}{(s-t)[s+\frac{1}{2}, k+1]} (n+\tfrac{1}{2}, 2s]\, \delta^{2s} u_0 \tag{5.6}$$

and the formula for the odd central differences of v_0 is

$$\mu\delta^{2t-1}v_0 = (-1)^{k-t}\frac{[t+\frac{1}{2},\, k]}{t!(k-t)!(n+\frac{1}{2},\, 2t]}\sum_{s=1}^{s=n}\frac{s(s-1)\dots(s-k)}{(s-t)[s+\frac{1}{2},\, k]}(n+\tfrac{1}{2},\, 2s]\,\delta^{2s-1}u_0. \tag{5.7}$$

where t=0, 1, 2, ..., k. The notation for the brackets are calculated as $[t+\frac{1}{2},\, k] = (t+\frac{1}{2})(t+\frac{1}{2}+1)\dots(t+\frac{1}{2}+k-1)/k!$ and $(n+\frac{1}{2},\, 2t] = (n-t+1)(n-t+2)\dots(n+t)/(2t)!$. Setting t=0 in Equation (5.6) gives Equation (5.5). Equation (5.6) is modified to get Equation (5.7). As mentioned, a table of the coefficients can be found in Sheppard's 1912 paper.

The smoothed values for the ends of the range are obtained by calculating the first smoothed value and its central differences from a specified set of u's. For example, if m=13 and j=2, a table for the central values of v_1 to v_7 can be constructed using v_7 and its first and second central differences. Calculate $v_6 = v_7 - \delta v_7$ then set the second central difference of v_6 equal to the second central difference of v_7. Calculate the first central difference of v_6 by $\delta v_7 + \delta^2 v_6$. Repeat the process to find v_5 to v_1. The last six smoothed values in the dataset are calculated in a similar manner.

To illustrate Sheppard's smoothing method using central differences, the first 13 values of the infant mortality dataset found in Appendix C will be used. The modern definition for mortality rate used by actuaries and scientists is described as the death rate or the number of deaths scaled to the size of a population per unit of time. The dataset gives the number of infant deaths under the age of one for every 1000 live births for 42 years and will be referred to as the infant dataset. We will use m=2n+1=13 and degree j=2. The central differences of u_7 are shown in Table 5.2.

Table 5.2: Central differences of u_7 using the infant dataset.

x	u	δu	$\delta^2 u$	$\delta^3 u$	$\delta^4 u$	$\delta^5 u$	$\delta^6 u$	$\delta^7 u$	$\delta^8 u$	$\delta^9 u$	$\delta^{10} u$	$\delta^{11} u$	$\delta^{12} u$
1870	137												
		0											
1871	137		-6										
		-6		12									
1872	131		6		-16								
		0		-4		21							
1873	131		2		5		-45						
		2		1		-24		127					
1874	133		3		-19		82		-321				
		5		-18		58		-194		665			
1875	138		-15		39		-112		344		-1104		
		-10		21		-54		150		-439		1323	
1876	**128**		**6**		**-15**		**38**		**-95**		**219**		**-386**
		-4		6		-16		55		-220		937	
1877	124		12		-31		93		-315		1156		
		8		-25		77		-260		936			
1878	132		-13		46		-167		621				
		-5		21		-90		361					
1879	127		8		-44		194						
		3		-23		104							
1880	130		-15		60								
		-12		37									
1881	118		22										
		10											
1882	128												

Taking u_7 and the even central differences (shown in bold) from Table 5.2, the smoothed value is

$$\begin{aligned} v_7 &= c_1 u_7 + c_2 \delta^2 u_7 + c_3 \delta^4 u_7 + c_4 \delta^6 u_7 + c_5 \delta^8 u_7 + c_6 \delta^{10} u_7 + c_7 \delta^{12} u_7 \\ &= 1(128) + 0(6) - 6(-15) - 8(38) - \frac{45}{11}(-95) - \frac{12}{13}(219) - \frac{1}{13}(-386) \\ &= 130.2 \end{aligned} \tag{5.8}$$

The next smoothed value can be obtained by extending Table 5.2 to u_8.

The first six smoothed values are obtained by calculating the first and second central differences of v_7 using Equations (5.6) and (5.7). Using Table 5.2 and the

coefficients from Sheppard's paper we calculate the first and second differences by

$$\begin{aligned}\delta v_7 &= \tfrac{1}{2}(c_1u_7 + c_2\delta u_7 + c_3\delta^3 u_7 + c_4\delta^5 u_7 + c_5\delta^7 u_7 + c_6\delta^9 u_7 + c_7\delta^{11}u_7)\\ &= \tfrac{1}{2}(1(-14) + 4(27) + \frac{36}{7}(-70) + \frac{20}{7}(205) + \frac{5}{7}(-659) + \frac{6}{91}(2260))\\ &= -0.99451 \end{aligned} \tag{5.9}$$

$$\begin{aligned}\delta^2 v_7 &= c_1u_7 + c_2\delta^2 u_7 + c_3\delta^4 u_7 + c_4\delta^6 u_7 + c_5\delta^8 u_7 + c_6\delta^{10} u_7 + c_7\delta^{12}u_7\\ &= 0(128) + 1(6) + \frac{20}{7}(-15) + \frac{20}{7}(38) + \frac{100}{77}(-95) + \frac{25}{91}(219) + \frac{2}{91}(-386)\\ &= 0.01898 \end{aligned} \tag{5.10}$$

The smoothed value and its first and second central differences from Equations (5.8), (5.9), and (5.10) are shown in bold in Table 5.3. The table can be constructed with these values by first obtaining the smoothed value, v_6, by $v_7 - \delta v_7 = 130.2 - (-0.99451) = 131.2$. Set the second central difference of v_6 to equal to the second central difference of v_7 to find the first central difference of v_6. The first central difference of v_6 is $\delta v_7 + \delta^2 v_6 = -0.99451 - 0.01898 = -1.01349$. Repeat the process to find the smoothed values v_5 to v_1.

5.3 Sheppard's Smoothing Formula in Terms of Central Summations

Sheppard developed a general solution to the problem in terms of central summations. He based his method on the fact that least squares gives the same result as the method of moments. The unknown constants in the polynomial can be expressed in terms of the moments, and the moments in turn can be expressed in terms of successive sums. It follows that the constants can be expressed in terms of summations. The alternate computational method follows the same principle and produce the same results as the

Table 5.3: Central differences of v_i.

i	v_i	δv_{i+1}	$\delta^2 v_i$
1	136.5		0.01898
		−1.08941	
2	135.4		0.01898
		−1.07043	
3	134.3		0.01898
		−1.05145	
4	133.2		0.01898
		−1.03247	
5	132.2		0.01898
		−1.01349	
6	131.2		0.01898
		−0.99451	
7	**130.2**		**0.01898**

method in terms of central differences but requires significantly less calculations. The advantage of using central summations is that the sums can be calculated successively in order to obtain the smoothed values.

The general form to calculate central sums are shown in Table 5.4.

Table 5.4: General form for calculating central sums.

i	u_i	Σ^1	Σ^2	Σ^3
1	u_1	u_1	u_1	u_1
2	u_2	u_1+u_2	$2u_1+u_2$	$3u_1+u_2$
3	u_3	$u_1+u_2+u_3$	$3u_1+2u_2+u_3$	$6u_1+3u_2+u_3$
⋮	⋮	⋮	⋮	⋮
n	u_n	Σu_i	$\Sigma(\Sigma_i^1)$	$\Sigma(\Sigma_i^2)$

Sheppard transforms the formula for the even central differences (5.6) into a formula involving successive sums. The formula can be found on page 104 of his 1914b paper. Setting t=0 the smoothing formula in terms of central summations for degree j=2 or 3 is

$$v_i = \mathrm{A}(\Sigma^1_{i+h} - \Sigma^1_{i-h-1}) + \mathrm{B}(\Sigma^2_{i+h-1} + \Sigma^2_{i-h+1}) + \mathrm{C}(\Sigma^3_{i+h-2} - \Sigma^3_{i-h+1}) \qquad (5.11)$$

where A, B, and C are coefficients and h is the bandwidth. Similar to the smoothing formulas using central differences, Sheppard derived the coefficients for the central sums and included them in his paper (Sheppard 1914b, p. 181). For example, if m=13 and j=2, the coefficients are A=$-\frac{11}{143}$, B=$\frac{11}{143}$ and C=$-\frac{2}{143}$. An extended formula is given for higher degrees of j.

All of the central smoothed values can be obtained using Equation 5.11. Alternatively, for a large dataset, the first three central smoothed values can be calculated using Equation 5.11 and then the third central difference can be obtained using a formula given by Sheppard on page 177 of his 1914b paper. Using the third central difference we can obtain the second and first central differences and the next smoothed value. The process is repeated to find the remaining central smoothed values.

The formula in Equation (5.11) gives all the central smoothed values in the dataset. A similar method is used to obtain the smoothed values corresponding to the bandwidth at the ends of the range. Taking the first set of m values, a summation table is constructed such that the successive sums are calculated upwards rather than downwards. Three values from the table are used to calculate the first smoothed value and its first and second central differences using the following three formulas:

$$v_1 = \alpha_1\Sigma_1^1 - \alpha_2\Sigma_2^2 + \alpha_3\Sigma_3^3 \tag{5.12}$$

$$\delta v_1 = -\beta_1\Sigma_1^1 + \beta_2\Sigma_2^2 - \beta_3\Sigma_3^3 \tag{5.13}$$

$$\delta^2 v_1 = \gamma_1\Sigma_1^1 - \gamma_2\Sigma_2^2 + \gamma_3\Sigma_3^3 \tag{5.14}$$

The coefficients, α_i, β_i, and γ_i, are found using formulas on page 156 of Sheppard's 1915 paper. We can then find v_2, v_3, ..., v_6 using v_1 and its first and second central differences in a similar manner to how we obtained v_6, v_5, ..., v_1 using v_7 and its first and second differences, shown in Table 5.3. For the last set of m values, the summation table is constructed such that the successive sums are calculated downwards as shown

in Table 5.4. The last smoothed value and its first and second central differences are obtained and the remaining smoothed values are calculated in a similar manner.

To illustrate the method in terms of central sums we will use the first 15 observations from the same infant dataset mentioned above. Taking the first 15 values of u, we can obtain three central smoothed values. The central sums for the first 15 values of u are shown in Table 5.5

Table 5.5: Central sums using the infant dataset.

i	u_i	Σ^1	Σ^2	Σ^3
1	137	**137**	**137**	**137**
2	137	**274**	**411**	**548**
3	131	405	816	1364
4	131	536	1352	2716
5	133	669	2021	4737
6	138	807	2828	7565
7	128	935	3763	11328
8	124	1059	4822	16150
9	132	1191	6013	22163
10	127	1318	7331	29494
11	130	1448	8779	**38273**
12	118	1566	**10345**	**48618**
13	128	**1694**	**12039**	**60657**
14	125	**1819**	**13858**	74515
15	126	**1945**	15803	90318

Taking the values shown in bold in Table 5.5, we calculate the following three smoothed values using Equation (5.11):

$$\begin{aligned} v_7 &= \mathrm{A}(\Sigma^1_{13} - \Sigma^1_0) + \mathrm{B}(\Sigma^2_{12} + \Sigma^2_0) + \mathrm{C}(\Sigma^3_{11} - \Sigma^3_0) \\ &= -\frac{11}{143}(1694 - 0) + \frac{11}{143}(10345 + 0) - \frac{2}{143}(38273 - 0) \\ &= 130.2 \end{aligned} \tag{5.15}$$

$$\begin{aligned} v_8 &= \mathrm{A}(\Sigma^1_{14} - \Sigma^1_1) + \mathrm{B}(\Sigma^2_{13} + \Sigma^2_1) + \mathrm{C}(\Sigma^3_{12} - \Sigma^3_1) \\ &= -\frac{11}{143}(1819 - 137) + \frac{11}{143}(12039 + 137) - \frac{2}{143}(48618 - 137) \\ &= 129.2 \end{aligned} \tag{5.16}$$

$$\begin{aligned} v_9 &= \mathrm{A}(\Sigma^1_{15} - \Sigma^1_2) + \mathrm{B}(\Sigma^2_{14} + \Sigma^2_2) + \mathrm{C}(\Sigma^3_{13} - \Sigma^3_2) \\ &= -\frac{11}{143}(1945 - 274) + \frac{11}{143}(13858 + 411) - \frac{2}{143}(60657 - 548) \\ &= 128.4 \end{aligned} \tag{5.17}$$

Table 5.5 is extended to obtain the remaining smoothed values, v_{10} to v_{36}. The same set of coefficients are used for each smoothed value. As mentioned, an alternate way for large datasets is to find the central differences of v_7, v_8, and v_9 to obtain the next smoothed value. The process is repeated to find the remaining central smoothed values.

To find the smoothed values for the first six u's we construct a table of sums using the first 13 u's. The sums are shown in Table 5.6.

Table 5.6: Central sums to obtain v_1.

i	u_i	Σ^1	Σ^2	Σ^3
1	137	**1694**	11677	57942
2	137	1557	**9983**	36282
3	131	1420	8426	**36282**
4	131	1289	7006	27856
5	131	1158	5717	20850
6	138	1025	4559	15133
7	128	887	3534	10574
8	124	759	2647	7040
9	132	635	1888	4393
10	127	503	1253	2505
11	130	376	750	1252
12	118	246	374	502
13	128	128	128	128

The values shown in bold in Table 5.6 are used in the formula for obtaining v_1. The first smoothed value and its first and second differences are calculated by the following three equations:

$$\begin{aligned} v_1 &= \alpha_1\Sigma_1^1 - \alpha_2\Sigma_2^2 + \alpha_3\Sigma_3^3 \\ &= \frac{517}{1001}(1694) - \frac{154}{1001}(9983) + \frac{22}{1001}(36282) \\ &= 136.5 \end{aligned} \tag{5.18}$$

$$\begin{aligned} \delta v_1 &= -\beta_1\Sigma_1^1 + \beta_2\Sigma_2^2 - \beta_3\Sigma_3^3 \\ &= -\frac{154}{1001}(1694) - \frac{66}{1001}(9983) - \frac{11}{1001}(36282) \\ &= -1.0989 \end{aligned} \tag{5.19}$$

$$\begin{aligned} \delta^2 v_1 &= \gamma_1\Sigma_1^1 - \gamma_2\Sigma_2^2 + \gamma_3\Sigma_3^3 \\ &= \frac{22}{1001}(1694) - \frac{11}{1001}(9983) + \frac{2}{1001}(36282) \\ &= 0.01898 \end{aligned} \tag{5.20}$$

Taking v_1 and its central differences we can calculate the smoothed values v_2 to v_6 in a similar manner as Table 5.3. As mentiond, the last six smoothed values of the dataset are obtained in a similar manner.

5.4 Sheppard's Smoothing Method Based on the Method of Least Squares

As mentioned, Sheppard's smoothed values can be obtained with successive application of ordinary least squares. The method of least squares considers an odd number of equally-spaced x values corresponding to y values. A polynomial, such as $y = a_0 + a_1x + a_2x^2$, is fitted to the data. The smoothed value is obtained for the central value by evaluating the polynomial at x=0. Moving one step to the right of the dataset, the process is repeated to obtain the next smoothed value. The smoothed values for the ends of the range are obtained by using the fitted values from the first and last polynomials.

Sheppard illustrated his method using central sums using the infant dataset (Sheppard 1914b). For comparison, the smoothing method was reproduced in R using least squares. Using m=13 points and a moving quadratic polynomial fitted by method of least squares the results were the same as Sheppard's. The resulting smoothed values (open circles) are shown in Figure 5.1.

Sheppard's method obtains the smoothed values but they do not visually lie on a 'smooth' curve. However, Sheppard wanted to find the best solution based on the mean square error of the smoothed values. Sheppard states that his smoothing methods in terms of central differences and central summations are equivalent to using the method of moments or least squares. Due to the amount of calculations required he informs the reader that the simplest method to use is central summations. Throughout his papers, Sheppard compares his smoothing method to other methods available at the time such as Spencer's graduation formula (Spencer 1904). He concludes that his method is the "best method" by using the smallest mean square error of the smoothed values as a criterion.

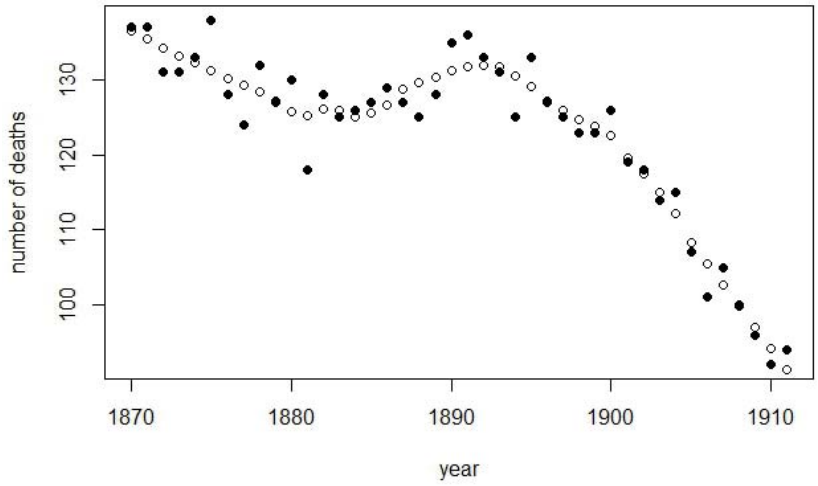

Figure 5.1: Sheppard's smoothed values (open circles) and the data (solid circles) using method of least squares.

5.5 Precursor Methods to Local Polynomial Regression

Sheppard's smoothing method was a precursor to local polynomial regression using a uniform kernel. Another precursor to local polynomial regression is Robert Henderson's smoothing method developed in 1916. Henderson developed a weighted local cubic fitting method. If w_h is the weight function for $h = -m,\ldots,m$, then the local cubic fit at i is

$$\sum_{h=-m}^{m} f(h) w_h u_{i+h} \tag{5.21}$$

where u_{i+h} are the observed number of deaths and $f(h)$ is a cubic polynomial whose coefficients have the property that the smoother reproduces the data if they are cubic. If W_h is symmetric then f is quadratic.

Frederick Macaulay describes E.T. Whittaker (1923) and along with Henderson (1924) smoothing methods in his 1931 book, *The Smoothing of Time Series*. The

Whittaker–Henderson graduation method is based on the minimization of an objective function:

$$f(u_1, u_2, \ldots, u_n) = \sum_{h=1}^{n} w_h(u_h - v_h)^2 + \lambda \sum_{h=1}^{n-k} (\Delta^k u_h)^2 \tag{5.22}$$

where λ is a constant parameter, w_1, w_2,...,w_n are the weights attributed to the squared deviations between the observed and graduated values, and $\Delta^k u_h$ is the kth forward difference of u_h and defined as:

$$\Delta^k u_h = \sum_{i=0}^{k} (-1)^i \binom{k}{i} u_{h+k-i} \tag{5.23}$$

The weights can be chosen such that less importance is placed on the number of deaths for the older ages where there are fewer individuals alive. In this case, the weights can be chosen to be inversely proportional to the estimated variance of the observed number of deaths. In his book, Macaulay shows how the the Whittaker–Henderson method can be used for time series models.

5.6 Comparing Sheppard's Methods to Modern Methods

5.6.1 Local Polynomial Regression

Sheppard's smoothing method is similar to local polynomial regression using a uniform kernel. Local polynomial regression is based on estimating $g(x_0)$ for any value x_0, using data in the immediate neighbourhood of x_0. Using a Taylor series about x_0, we have

$$g(x) = g(x_0) + (x - x_0)g^{(1)}(x_0) + \tfrac{1}{2}(x - x_0)^2 g^{(2)}(x_0) + \ldots \tag{5.24}$$

Let $\beta_j = g^{(j)}(x_0)/j!$, for $j = 0,1,\ldots,p$, then a local approximation to $g(x)$ is

$$g(x) \doteq \sum_{j=0}^{p} (x - x_0)^j \beta_j. \tag{5.25}$$

Local polynomial regression is weighted least squares where the weights are taken to be high for x values close to x_0 and lower for x values further away from x_0. Sheppard used a uniform kernel for choosing the weights:

$$W_h(x) = \begin{cases} \frac{1}{2h}, & \text{if } -h < x < h \\ 0, & \text{otherwise} \end{cases} \tag{5.26}$$

The parameter h is called the bandwidth. $W_h(x - x_0)$ takes on the value $\frac{1}{2h}$ for all x values within a radius of h units of x_0, and it takes on the value 0 for all x values further away from x_0. The weighted least squares problem is to minimize

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} (x_i - x_0)^j \beta_j \right)^2 W_h(x_i - x_0) \tag{5.27}$$

with respect to β_0, β_1, ..., β_p.

Sheppard used a quadratic polynomial with m=13 values to illustrate his method using the infant death dataset. To compare his method to local polynomial regression with a uniform kernel, a bandwidth h=6 was used. The differences between Sheppard's smoothed values and the smoothed values obtained using local polynomial regression are given in Figure 5.2. The values are identical except for the first and last six values representing half the bandwidth. Sheppard used the first and last polynomials to find the ends of the range, whereas local polynomial regression uses an asymmetrical bandwidth to obtain the ends of the range. For example, the sixth value is smoothed by using the first 12 values (5 to the left and 6 to the right) rather than 13 and evaluating the polynomial at x_6.

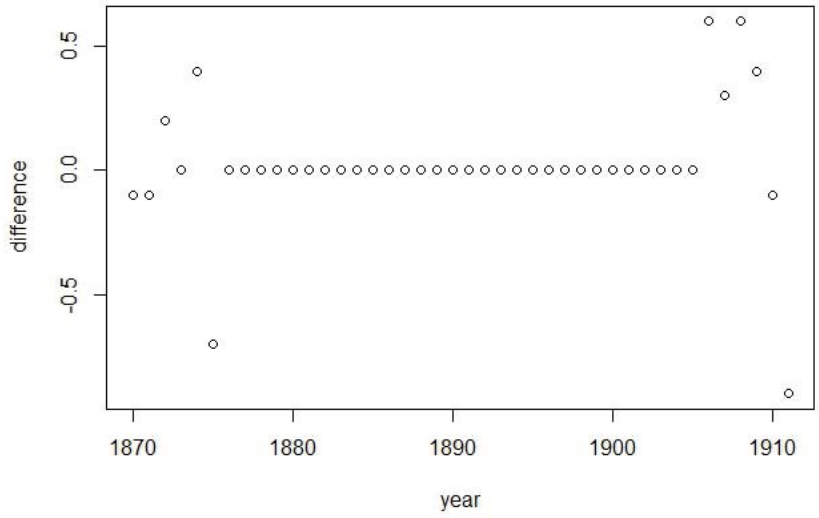

Figure 5.2: Differences between the smoothed values using Sheppard's method and local polynomial regression.

5.6.2 Bayesian Smoothing Method

Sheppard's smoothing method was compared to modern Bayesian smoothing. Since the dataset consists of the number of infant deaths per 1000 live births spanning 42 years from 1870 to 1911, a continuous smoothing function is desired to represent the rate of deaths per year. The function is modelled by

$$\mu(t) = \sum_{i=1}^{p} b_i(t)\beta_i = \mathbf{B}\boldsymbol{\beta} \tag{5.28}$$

where t is the year, β_i for $i = 1 \dots p$ are unknown parameters, $b_i(t)$ are basis functions and p is the number of basis functions. In vector notation, $\mathbf{B}$ is an $n \times p$ matrix and $\boldsymbol{\beta}$ is a $p \times 1$ vector. A cubic B-spline basis was chosen as the basis for this model. The dimension of the B-spline basis is the number of knots plus the degree of the polynomial plus an intercept. The location of the knots can be chosen to be at the predictor variables or any other suitable location depending on the model. For our

model, they are chosen to be at the predictor variables from 1870 to 1911. More details on the B-spline basis can be found in §6.3. We define the total proportion of deaths for each year as

$$\mu_j = \int_{t_j}^{t_{j+1}} \mu(t)\, dt \tag{5.29}$$

We assume the data which consists of the proportion of deaths for each year follows a normal distribution. This gives the following likelihood function:

$$g(y_1, \ldots, y_n | \beta_1, \ldots, \beta_n) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_j - \mu(t_j))^2 \right) \tag{5.30}$$

where y_j is the proportion of deaths per year, μ_j is the mean death rate for each year and n is the number of years. A possible prior distribution is given by

$$f(\beta_1, \ldots, \beta_n) \propto \exp\left(-\frac{k}{2} \int_0^{\infty} (\mu''(t))^2 dt \right) \tag{5.31}$$

where k is a constant. The exponent in the prior distribution is known as a smoothing function or penalty function in other contexts. These functions penalize the roughness of a curve. As $k \to 0$ the prior allows $\mu(t)$ to be less smooth and as $k \to \infty$ the prior forces $\mu''(t) = 0$, ie. $\mu(t)$ tends to a straight line. For our parameterization, the prior distribution is an improper Gaussian density in the parameters $\boldsymbol{\beta}$. The prior distribution can be written as:

$$f(\beta_1, \ldots, \beta_n) \propto \exp\left(-\frac{k}{2} \int_0^{\infty} (\mu''(t))^2 dt \right) = -\frac{k}{2} \boldsymbol{\beta}^{\boldsymbol{t}} \mathbf{S} \boldsymbol{\beta} \tag{5.32}$$

where $\mathbf{S}$ is a $p \times p$ matrix. The derivation of the solution of the integral is shown in §6.3. Ignoring the constant $\frac{1}{\sqrt{2\pi\sigma^2}}$ in the likelihood (5.30), since it does not depend on the parameters, the posterior distribution with unknown mean and known variance

becomes

$$f(\beta_1,\ldots,\beta_n|y_1,\ldots,y_n) \propto \prod_{j=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_j-\mu(t_j))^2\right)\exp\left(-\frac{k}{2}\int_0^\infty (\mu''(t))^2 dt\right)$$
$$= \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\mathbf{B}\boldsymbol{\beta})^{\boldsymbol{t}}(\boldsymbol{y}-\mathbf{B}\boldsymbol{\beta})\right) + \exp\left(-\frac{k}{2}\boldsymbol{\beta}^{\boldsymbol{t}}\mathbf{S}\boldsymbol{\beta}\right) \tag{5.33}$$

The exponent in the posterior distribution can be simplified as

$$-\frac{1}{2}\left(\sigma^{-2}(\boldsymbol{y}-\mathbf{B}\boldsymbol{\beta})^{\boldsymbol{t}}(\boldsymbol{y}-\mathbf{B}\boldsymbol{\beta}) + k\boldsymbol{\beta}^{\boldsymbol{t}}\mathbf{S}\boldsymbol{\beta}\right)$$
$$= -\frac{1}{2}\left(\sigma^{-2}(\boldsymbol{y}^{\mathbf{t}}\boldsymbol{y} - 2\boldsymbol{\beta}^{\boldsymbol{t}}\mathbf{B}^{\boldsymbol{t}}\boldsymbol{y} + \boldsymbol{\beta}^{\boldsymbol{t}}\mathbf{B}^{\boldsymbol{t}}\mathbf{B}\boldsymbol{\beta}) + k\boldsymbol{\beta}^{\boldsymbol{t}}\mathbf{S}\boldsymbol{\beta}\right)$$
$$= -\frac{1}{2}\left(\boldsymbol{\beta}^{\boldsymbol{t}}\left(\frac{\mathbf{B}^{\boldsymbol{t}}\mathbf{B}}{\sigma^2}+k\mathbf{S}\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\boldsymbol{t}}\frac{\mathbf{B}^{\boldsymbol{t}}\boldsymbol{y}}{\sigma^2} + \frac{\boldsymbol{y}^{\mathbf{t}}\boldsymbol{y}}{\sigma^2}\right) \tag{5.34}$$

Completing the square centered around

$$\boldsymbol{\theta} = \left(\frac{\mathbf{B}^{\boldsymbol{t}}\mathbf{B}}{\sigma^2}+k\mathbf{S}\right)^{-1}\frac{\mathbf{B}^{\boldsymbol{t}}\boldsymbol{y}}{\sigma^2} \tag{5.35}$$

gives

$$-\frac{1}{2}\left(\boldsymbol{\beta}^{\boldsymbol{t}}\left(\frac{\mathbf{B}^{\boldsymbol{t}}\mathbf{B}}{\sigma^2}+k\mathbf{S}\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\boldsymbol{t}}\left(\frac{\mathbf{B}^{\boldsymbol{t}}\mathbf{B}}{\sigma^2}+k\mathbf{S}\right)\boldsymbol{\theta} + \boldsymbol{\theta}^{\boldsymbol{t}}\left(\frac{\mathbf{B}^{\boldsymbol{t}}\mathbf{B}}{\sigma^2}+k\mathbf{S}\right)\boldsymbol{\theta} - \boldsymbol{\theta}^{\boldsymbol{t}}\left(\frac{\mathbf{B}^{\boldsymbol{t}}\mathbf{B}}{\sigma^2}+k\mathbf{S}\right)\boldsymbol{\theta} + \frac{\boldsymbol{y}^{\mathbf{t}}\boldsymbol{y}}{\sigma^2}\right)$$
$$= -\frac{1}{2}\left((\boldsymbol{\beta}-\boldsymbol{\theta})^{\boldsymbol{t}}\left(\frac{\mathbf{B}^{\boldsymbol{t}}\mathbf{B}}{\sigma^2}+k\mathbf{S}\right)(\boldsymbol{\beta}-\boldsymbol{\theta}) - \boldsymbol{\theta}^{\boldsymbol{t}}\left(\frac{\mathbf{B}^{\boldsymbol{t}}\mathbf{B}}{\sigma^2}+k\mathbf{S}\right)\boldsymbol{\theta} + \frac{\boldsymbol{y}^{\mathbf{t}}\boldsymbol{y}}{\sigma^2}\right) \tag{5.36}$$

Ignoring the last two terms, since they do not depend on $\boldsymbol{\beta}$, the posterior distribution is

$$f(\boldsymbol{\beta}|\boldsymbol{y}) \propto \exp\left[-\frac{1}{2}\left((\boldsymbol{\beta}-\boldsymbol{\theta})^{\boldsymbol{t}}\left(\frac{\mathbf{B}^{\boldsymbol{t}}\mathbf{B}}{\sigma^2}+k\mathbf{S}\right)(\boldsymbol{\beta}-\boldsymbol{\theta})\right)\right]. \tag{5.37}$$

The posterior is a multivariate normal distribution where the mean of $\boldsymbol{\beta}$ has expected value $\boldsymbol{\theta} = (\frac{\mathbf{B}^{\boldsymbol{t}}\mathbf{B}}{\sigma^2}+k\mathbf{S})^{-1}\frac{\mathbf{B}^{\boldsymbol{t}}\boldsymbol{y}}{\sigma^2}$ and variance equal to $(\frac{\mathbf{B}^{\boldsymbol{t}}\mathbf{B}}{\sigma^2}+k\mathbf{S})^{-1}$.

Figures 5.3 to 5.6 show $\mu(t)$ using $k = 0.1$, 1, 5 and 10. The parameter σ^2 is assumed to be known and set to equal 1. As k increases the curve becomes more smooth. When $k = 0.1$ the curve is too wiggly and not what we would expect for a curve showing the number of infant deaths. When $k = 10$ the curve appears to be too smooth and underfits the data. When $k = 5$ the curve is what we would expect to find and does not underfit or overfit the data. Figure 5.10 shows the residual plot when $k = 5$.

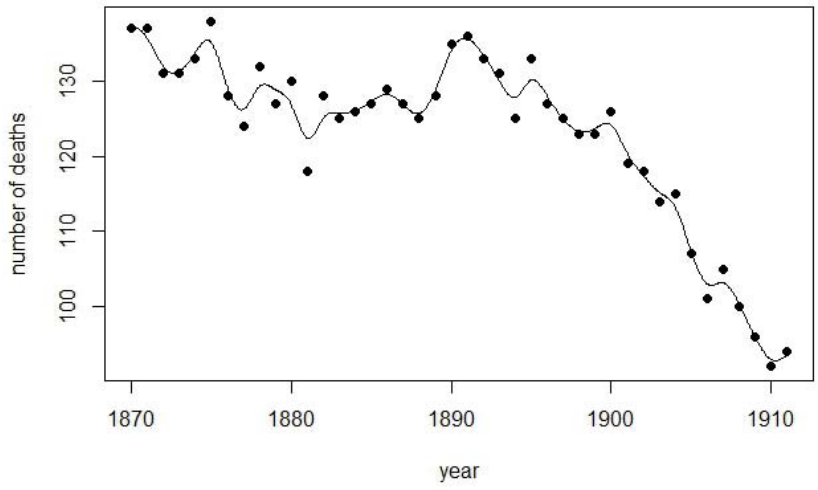

Figure 5.3: Bayesian smoothing model using k=0.1.

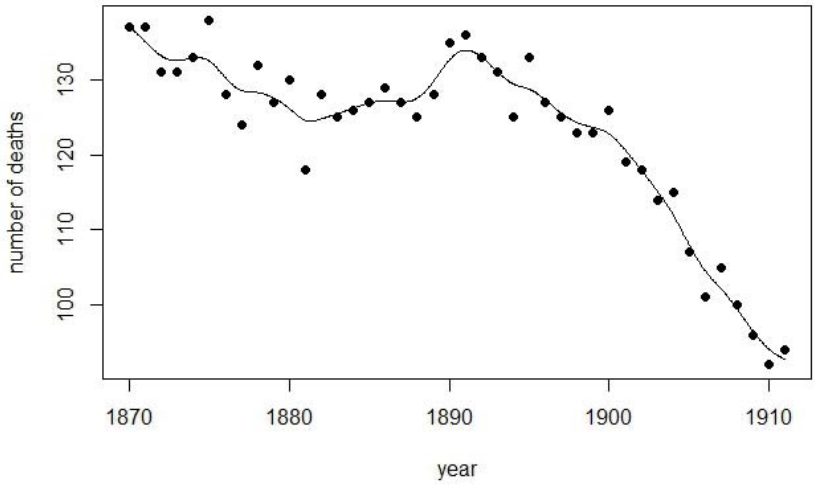

Figure 5.4: Bayesian smoothing model using k=1.

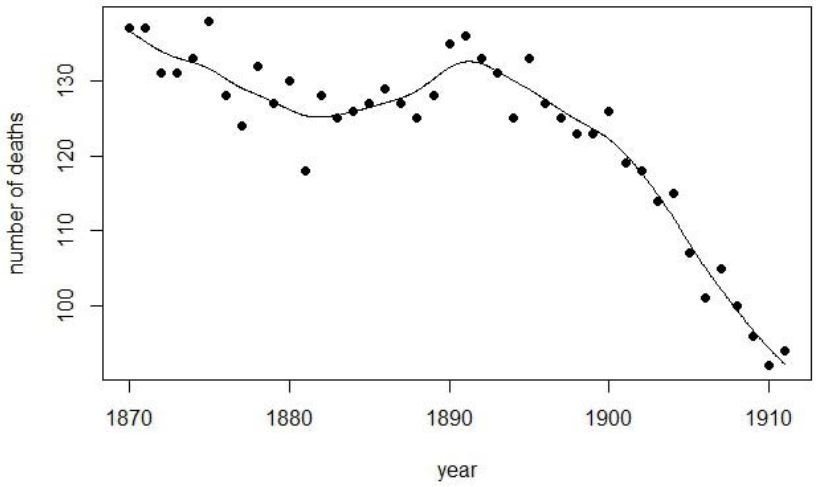

Figure 5.5: Bayesian smoothing model using k=5.

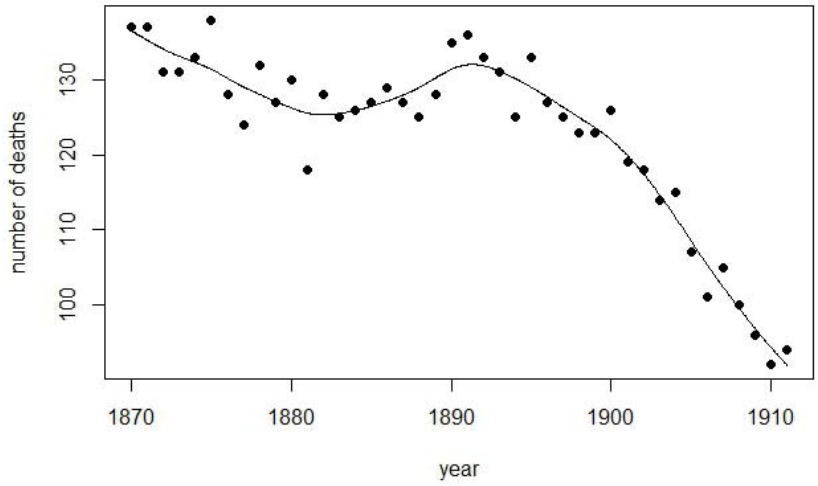

Figure 5.6: Bayesian smoothing model using k=10.

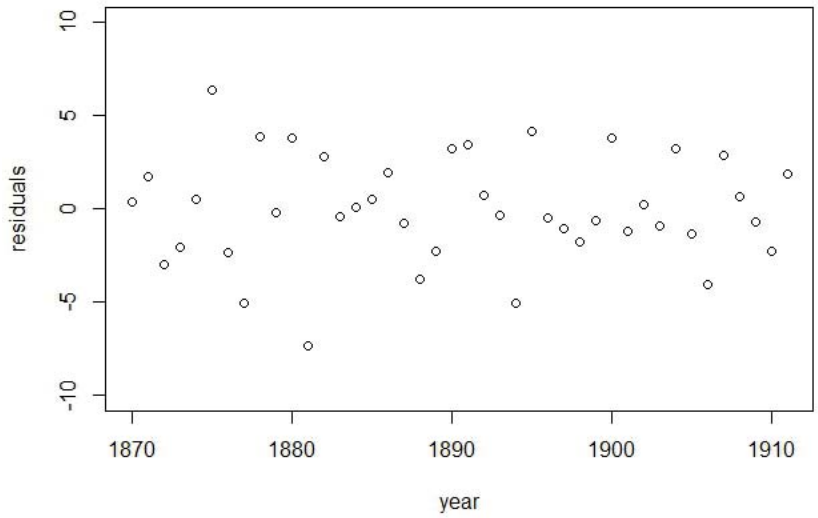

Figure 5.7: Residual plot using k=5.

Credible intervals are constructed using

$$\mathrm{E}(\mu(t_j)) = (\mathbf{B}\boldsymbol{\beta})_j = (\mathbf{B}\boldsymbol{\theta})_j \tag{5.38}$$

and

$$\mathrm{Var}(\mu(t_j)) = \mathrm{diag}(\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^{\boldsymbol{t}}) \tag{5.39}$$

where $\boldsymbol{\Sigma} = (\frac{\mathbf{B}^{\boldsymbol{t}}\mathbf{B}}{\sigma^2} + k\mathbf{S})^{-1}$ is the variance of $\boldsymbol{\beta}$. The 95% credible interval for $\mu(t_j)$ is

$$(\mathbf{B}\boldsymbol{\theta})_j \pm (1.96)\sqrt{\mathrm{diag}(\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^{\boldsymbol{t}})} \tag{5.40}$$

The 95% credible intervals are shown in Figure 5.8. There is a 95% probability that $\mu(t_j)$ lies between the upper and lower bands.

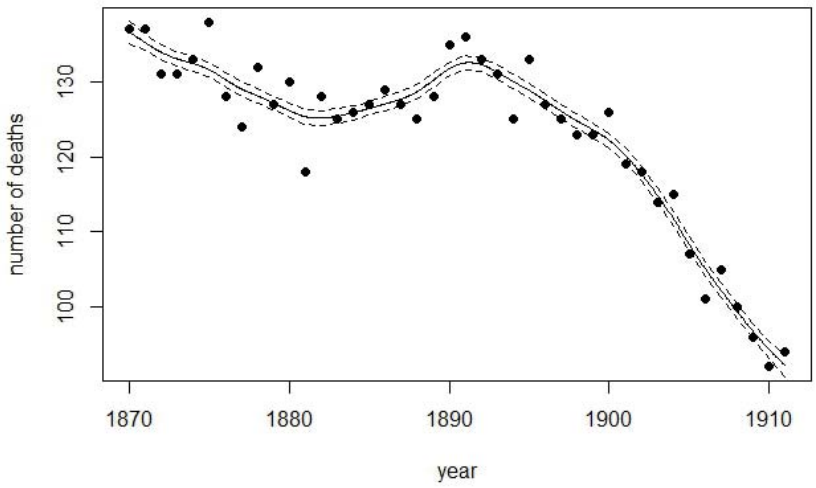

Figure 5.8: The 95% credible intervals using k=5.

Figure 5.9 compares Sheppard's smoothed values (open circles) and the Bayesian model with k=5. Figure 5.10 shows the differences between Sheppard's smoothed values and the Bayesian model $\mu(t)$ evaluated yearly.

The Bayesian model is similar to Sheppard's smoothed values. However, the Bayesian model fits the data more closely. The model takes the entire dataset into consideration whereas Sheppard's method uses a fixed bandwidth. This leads to Sheppard's values not being consistently smooth throughout the range of the data. The smoothed values for 1882 and 1883 are too high when compared to the smoothed values for the years on either side (1881 and 1884). There is a discontinuity between 1886 and 1887, while the smoothed values are linear before and after these years. The rest of Sheppard's smoothed values are similar to the Bayesian fit.

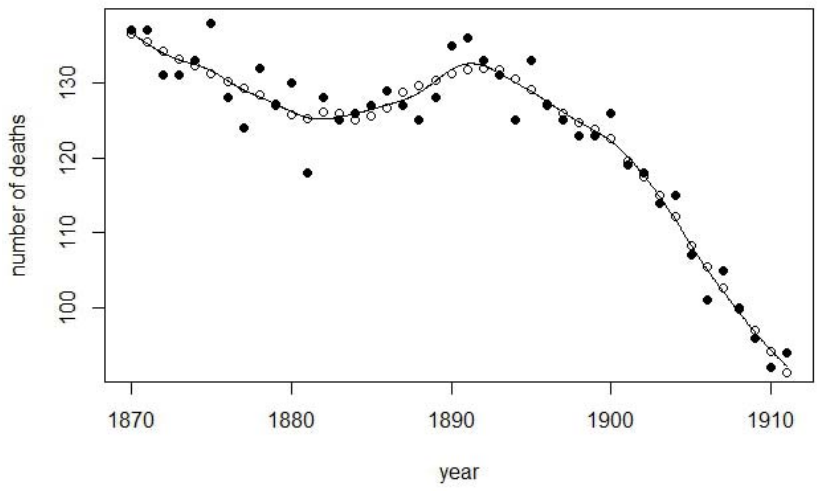

Figure 5.9: Comparison of Sheppard's smoothed values (open circles), Bayesian smoothing (line) using k=5 and the data (solid circles).

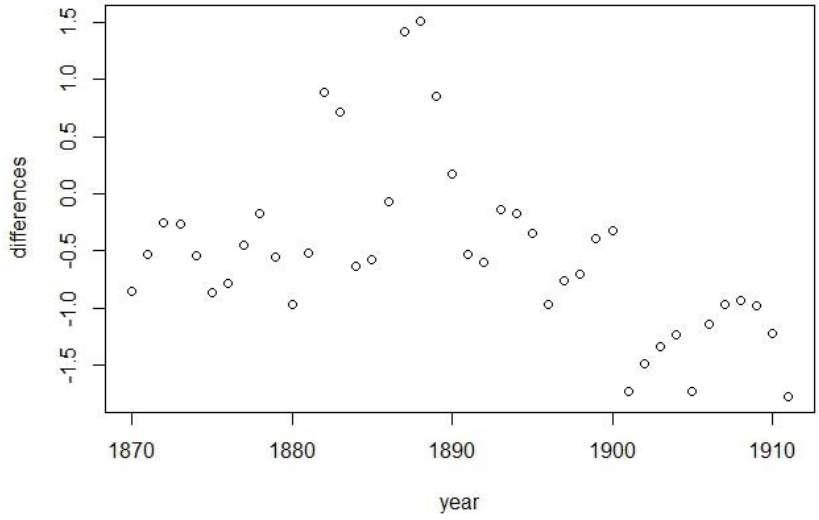

Figure 5.10: Differences between Sheppard's smoothed values and Bayesian $\mu(t)$ evaluated yearly.

5.7 Conclusion

Sheppard's smoothing method requires datasets with equally-spaced values of x. A disadvantage of his method is the problem at the boundaries. As we have seen, if $m = 13$, the first six and the last six values of the dataset are smoothed using the first and last fitted polynomials. This results in not having all of the values smoothed in the same way.

It was Sheppard's attempt with his smoothing method to find the best solution based on the mean square error of the smoothed values. If least squares is used to fit the entire dataset, the sum of the errors is zero but the curve is comprised of perturbations. On the other hand, if a perfectly smooth curve is fitted to the data, such as a parabola, the result will be a large mean square error. With regards to the process of smoothing, Sheppard writes,

But this is not a scientific method, unless some criterion is adopted for

> deciding which of the possible new sequences is the best; and it is difficult to apply such a criterion except to the sequence as a whole, in which case the process becomes one of "fitting." (Sheppard 1912)

The problem is to find a balance between smoothing and fitting. It is still the same issue as today where there is a trade-off between smoothing and bias.

Sheppard's smoothing methods were adopted by other scientists such as Catherine W.M. Sherriff (Sherriff 1920), geneticist Julia Bell (Rhodes 1921), and mathematician Oskar Anderson (Anderson 1927). Sherriff worked on graduation formulae by fitting higher order parabolas using Sheppard's method and wrote an article describing the relationship between Sheppard's and Spencer's graduation formulae. Her conclusions state that Sheppard's method removes errors more successfully than Spencer's formula since Sheppard's method is based on least squares. Oskar Anderson used Sheppard's method using central differences for time series models.

Chapter 6

The Development of Bayesian Smoothing

6.1 Background

This chapter provides an overview of the development of Bayesian smoothing.

In 1763 Richard Price presented a paper (Price and Bayes 1763) to the Royal Statistical Society showing the proof of Bayes' theorem. Price was the compiler of the Northampton life table discussed in Chapter 2. The paper was published in *Philosophical Transactions* posthumously after Thomas Bayes' death. Bayes was a mathematician, philosopher and Presbyterian minister. The paper begins with an introduction written by Price on the philosophical basis of Bayesian probability followed by an essay written by Bayes outlining his theorem. Mathematician Pierre Simon Laplace introduced the same theorem independently in 1774 in the *Mém. de l'Académie royale des sciences presentés par divers savans.* Laplace was working on a problem between 1774 and 1781 on birth ratios. Using the birth records in Paris spanning 26 years between 1745 to 1770, Laplace conducted a test to see if the probability of a birth being male was greater than 1/2. He calculated the probability

of a birth being male using a uniform prior as 0.50971. Using birth records from London spanning 93 years between 1664 to 1756, Laplace calculated the probability of a birth being male (using a uniform prior) as 0.51346. Based on the data from both Paris and London, the probability of there being more males in London than Paris given the data (still using a uniform prior) was $\frac{1}{410{,}458}$. Laplace (1781) concluded there was likely a probable cause, such as climate, food, or customs for London having a higher ratio of male births than Paris. This was the beginning of the Bayesian view.

6.2 The Bayesian View

In the Bayesian approach, we model uncertainty as a distribution of the parameters. The prior distribution represents our uncertainty before viewing the data. The posterior distribution (the conditional distribution of unknown parameters (unobserved data) given the observed data) represents our uncertainty after viewing the data. Bayes' theorem implies that the posterior distribution is proportional to the product of the prior distribution and the likelihood. The likelihood is the weight given to each of the unobservable events given the occurrence of the unknown parameters. The prior distribution is an initial estimate of the probability of the unknown parameters based on prior knowledge or experience. It is subjective since personal beliefs can vary from person to person. New evidence can change our beliefs, and thus Bayes' theorem allows for the model to be updated with revised probabilities. Bayesian statistics is predictive meaning we can find the conditional probability distribution of the next observation given the data. In contrast to the frequentist approach, we use the empirical distribution of the statistic over all the samples obtained rather than the sampling distribution over all possible repetitions.

As mentioned in Chapter 3, Bayes theorem was routinely used during the nineteenth century and was referred to as the Gaussian method of inverse probability (Stigler 2008, p. 5). Its use had diminished by the early twentieth century but inter-

est was renewed by mid-twentieth century by statisticians such as H. Jeffreys (1946), L.J. Savage (1954), B. De Finetti (1961, 1974), and D.V. Lindley (1965).

6.3 Bayesian Smoothing

The fundamentals of Bayesian smoothing were first introduced in 1970 by statisticians George Kimeldorf and Grace Wahba. Kimeldorf and Wahba (1970) explored the relationships between Bayesian estimation and spline smoothing in *A correspondence between Bayesian estimation on stochastic processes and smoothing by splines*. They proved that polynomial spline smoothing is equivalent to Bayesian estimation under a class of improper Gaussian prior distributions. Wahba extended the methodology in 1978 in *Improper priors, spline smoothing and the problem of guarding against model errors in regression*. She showed that spline and generalized spline smoothing is equivalent to Bayesian estimation with a partially improper Gaussian prior. The commonly used roughness penalty (quadratic) is equivalent to a partially improper Gaussian prior in the sense that the smoothing spline estimator can be interpreted as the mean of the corresponding Gaussian posterior. Wahba includes some computational tricks for the methods described in the paper.

The development of Bayesian smoothing went through a period of computational difficulty. The formulas presented by Kimeldorf and Wahba were not practical given the computing power available at the time. Statisticians paid little attention to Bayesian smoothing splines in the early 1970's and few papers were published by the mainstream statistical journals. By the late 1970's an increase in computing power and the implementation of simulation methods became available, and smoothing splines could be calculated for large datasets. Additionally, a good data-based method for choosing the smoothing paramater was found and multivariate smoothing methods were developed.

In 1990 Wahba published a book titled *Spline Models for Observational Data*

describing Bayesian smoothing in extensive detail. Topics include splines, partial splines, estimating the smoothing parameter, and Bayesian intervals. Wahba shows that all smoothing spline models have a Bayesian interpretation. The smoothing spline estimator is equivalent to the mean of the posterior. This allows for inferences to be made using Bayesian credible intervals. Wahba's book was published around the same time Markov Chain Monte Carlo (MCMC) methods were starting to be confidently accepted and used by mainstream statisticians. In general, MCMC methods are based on sampling from an approximate distribution and then correcting the samples to better approximate the target distribution (posterior distribution). Robert and Casella (2011) give an interesting short history of MCMC. These simulation methods are useful when direct sampling from the posterior distribution is difficult. As we have seen in Chapter 5, if a conjugate prior is used the posterior distribution can be easily determined since it is a closed form. If a conjugate prior is not available, as is the case for most Bayesian models, a computational sampling method is implemented to approximate the posterior distribution.

MCMC methods include the Metropolis (1953) and Metropolis-Hastings (1970) algorithms and the Gibbs sampler (Geman and Geman 1984). In general, the algorithms are used to draw samples (iteratively) from the target distribution and update the parameters. The Metropolis and Metropolis-Hastings algorithms require knowing the joint density function of the target distribution up to a constant of proportionality. The Gibbs sampler is a special case of the Metropolis-Hastings algorithm and requires knowing all of the conditional target distributions.

6.4 Bayesian Smoothing and Mortality Data

Bayesian smoothing has many applications. A selection of Bayesian smoothing or graduation methods using mortality data include Kimeldorf and Jones (1967), Hickman and Miller (1977), Cornfield and Detre (1977), Carlin (1992), Congdon (2009),

Luoma *et al.* (2012), and Dellaportas *et al.* (2001). Kimeldorf and Jones (1967) give a theoretical method of graduation based on what they refer to as 'personal probability' also known as Bayesian statistics. They include a numerical example to compare their method to Whittaker's method. Whittaker's method is algebraically similar but proceeds from a frequentist point of view regarding the parameters as fixed and does not take recent observations into account. Hickman and Miller (1977) review Kimeldorf and Jones's paper and attempt to develop a way for actuaries to deal with some of the technical issues they identified in their method. Cornfield and Detre (1977) derive the moments of the posterior probability distribution function. Carlin (1992) provides simple Bayesian models (non-parametric) for smoothing data using Monte Carlo techniques and the Gibbs sampler with the hope of making Bayesian smoothing more feasible for actuaries. Congdon (2009) used Bayesian smoothing to model life expectancy for 1,118 small areas in Eastern England over a five-year period 1999 to 2003. Luoma *et al.* (2012) proposed a two-dimensional (cohort) smoothing spline method using mortality data.

Dellaportas *et al.* (2001) show how simulation-based Bayesian smoothing can be used to construct life tables. Four advantages to using Bayesian inference over other methods are given: the parameters have a straightforward interpretation (the use of prior distributions avoids overparameterization), the non-normality of the likelihood means that the least square estimates are inadequate, application to incomplete life tables can use simulation-based computation, and quantities such as the joint lifetime of a couple or the median lifetime of a person can be derived from the posterior densities. A non-linear logistic Bayesian model, a model accounting for extra-binomial variation, and a log-normal Bayesian model are derived. They illustrate and compare the three models using mortality data from 1988 to 1992 of English and Welsh females which is defined as a complete life table. An incomplete or abridged life table is comprised of mortality data collected by age groups at five year intervals rather than individual years except for the first five years which are in two intervals, [0, 1) and [1, 5). Incomplete life tables are common in countries that do not collect and record

vital statistics adequately. The incomplete table is extended to a complete life table using an MCMC strategy by sampling from a model that is proportional to the full model used with the complete life table. The results are compared using boxplots of the posterior marginal distributions.

Simulation-based Bayesian smoothing has been extended to applications such as environmental studies, finance, actuarial science, spatial and biological statistics, physics and astronomy, and medicine. An adaptive MCMC method was introduced in 2001 by H. Haario, E. Saksman and J. Tamminen. Advances in MCMC methods involve adaptive Metropolis-Hastings random walk samplers and Metropolis within a Gibbs sampler. Innovations in MCMC methods continue to be developed to enhance Bayesian inference.

6.5 Conclusion

The Bayesian approach has a long history but the development of Bayesian smoothing was slow to evolve. The posterior distribution is seldom in closed form which requires computing power; the implementation of MCMC methods made many calculations feasible. Bayesian smoothing is used in a wide range of applications including modelling mortality data for the construction of life tables. Bayesian smoothing can be used to smooth out the irregularities of complete and incomplete life tables while taking into account past observations. Advances in simulation-based Bayesian smoothing are ongoing.

Chapter 7

Bayesian Smoothing

7.1 The Objective

The objective of the Bayesian model described in this chapter is to predict the probability of life using eighteenth-century mortality data and modern smoothing techniques, and compare the results to eighteenth-century smoothing methods.

7.2 Preliminary Analysis of the Data

The data comes from the Bills of Mortality recorded on a broadside held in the Guildhall Library, London, England (Smart 1738b). As mentioned in Chapter 2, starting in the early seventeenth century, the Bills of Mortality for the City of London were published weekly to warn residents of possible outbreaks of the bubonic plague. John Smart, a clerk at the Guildhall in London, was the first to compile the data and construct a life table (see Appendix A) in order to estimate annuities (Hald, 1990, p.518). More details about the table are discussed in Chapter 2.

The data gives the number of deaths for each year between 1728 to 1737 inclusive for each age group ranging from birth to greater than 90 years of age. Table 2.3 shows the aggregate data of the number of deaths for the decade corresponding to the twelve age groups. Smart made the convenient assumption that the population in London was stationary. Observing Smart's life table in Appendix A we find the yearly rates remain constant over a few years and conclude that the resulting smoothed values are piecewise linear. The cumulative number of deaths per thousand versus age as reported by Smart is shown in Figure 7.1. The solid circles denote the rates for each age group given in the dataset. The open circles denote Smart's calculations for the yearly rates. We observe that the plot (Figure 7.1) displays some curvature. This is because it is the cumulative distribution of Smart's life table which is piecewise linear.

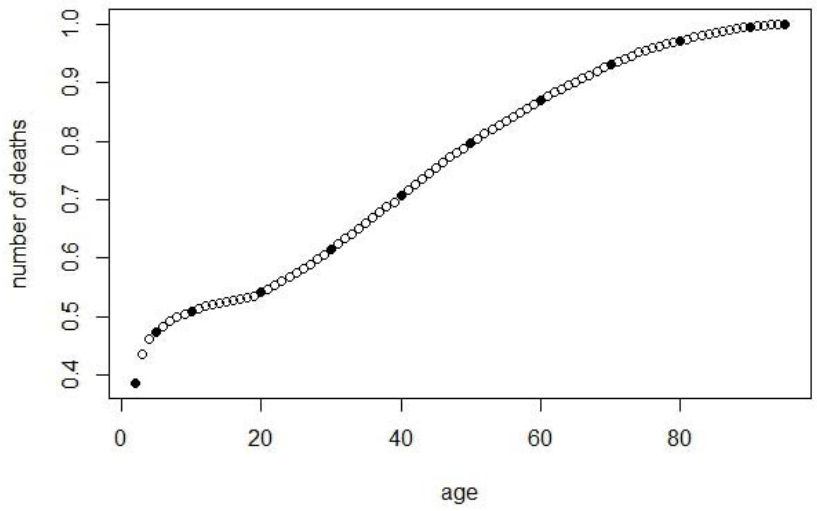

Figure 7.1: Cumulative number of deaths per thousand versus age as reported by Smart (open circles) and group data (solid circles).

7.3 The Model: Bayesian Smoothing

In general, a Bayesian posterior distribution is given by *posterior* $\propto$ *likelihood* $\times$ *prior*

$$f(a_1, \ldots, a_n | y_1, \ldots, y_n) \propto g(y_1, \ldots, y_n | a_1, \ldots, a_n) \times f(a_1, \ldots, a_n). \tag{7.1}$$

Our goal is to model the number of deaths for each year of life from birth to 100. Let $\lambda(t)$ be a continuous function that represents the instantaneous rate of death per year at age t. We assume deaths for each year follow a Poisson process. Since the aggregate data is the total number of deaths for the decade grouped by age and we want to predict the number of deaths for each year, a smooth continuous function is desired. The function is modelled by

$$\lambda(t) = \sum_{i=1}^{p} a_i b_i(t) \tag{7.2}$$

where t is the age in years, a_i for $i = 1 \ldots p$ are unknown parameters, $b_i(t)$ are basis functions and p is the number of basis functions.

A cubic B-spline basis is constructed using the built-in R function `bs()`. The dimension of the B-spline basis is the number of knots plus the degree of the polynomial plus an intercept. The location of the knots can be chosen to be at the predictor variables or any other suitable location depending on the model. Figure 7.2 shows cubic B-splines on the interval $[0, 100]$ at the age group boundaries. The knot locations have been highlighted using as upward ticks in the x-axis.

We define the total death rate for each age group as

$$\lambda_j = \int_{t_j}^{t_{j+1}} \lambda(t)\, dt \tag{7.3}$$

where each age group j consists of ages $t_j < t < t_{j+1}$.

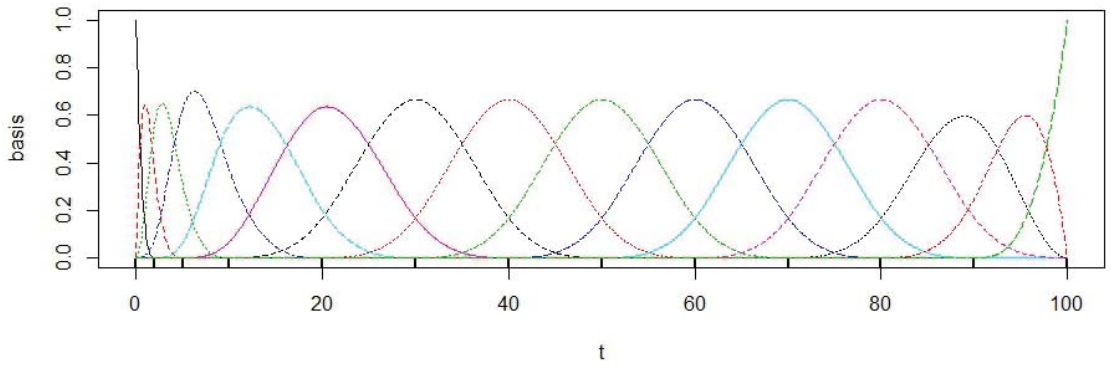

Figure 7.2: Cubic B-splines on $[0, 100]$ corresponding to knots at 0, 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100.

Because we assumed a Poisson process for deaths, the number of deaths follows a Poisson distribution in each age group cell. This gives the following likelihood function:

$$g(y_1, \ldots, y_n | a_1, \ldots, a_n) = \prod_{j=1}^{n} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!} \tag{7.4}$$

where y_j is the death count for each age group, λ_j is the total death rate for each age group and n is the number of groups. A possible prior distribution is given by

$$f(a_1, \ldots, a_n) \propto \exp\Big(- k \int_0^{\infty} (\lambda''(t))^2 dt \Big) \tag{7.5}$$

where k is a constant. The exponent in the prior distribution is known as a smoothing function or penalty function in other contexts. These functions penalize the roughness of a curve. As $k \to 0$ the prior allows $\lambda(t)$ to be less smooth and as $k \to \infty$ the prior forces $\lambda''(t) = 0$, ie. $\lambda(t)$ tends to a straight line. For our parameterization, the prior distribution is an improper Gaussian density in the parameters $\boldsymbol{a}$. The posterior becomes

$$f(a_1, \ldots, a_n | y_1, \ldots, y_n) \propto \Big(\prod_{j=1}^{n} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!} \Big) \exp\Big(- k \int_0^{\infty} (\lambda''(t))^2 dt \Big) \tag{7.6}$$

Taking the natural log and ignoring the constant $y_j!$, since it does not depend on the parameters, the log posterior distribution becomes

$$\log f(a_1, \ldots, a_n | y_1, \ldots, y_n) = \sum_{j=1}^{n} (y_j \log(\lambda_j) - \lambda_j) - k \int_0^\infty (\lambda''(t))^2 dt. \tag{7.7}$$

where the log of the constant of proportionality has been suppressed. The log prior distribution can be written as

$$\log f(\boldsymbol{a}) = -k \int_0^\infty (\lambda''(t))^2 dt = -k \boldsymbol{a}^{\boldsymbol{t}} \mathbf{S} \boldsymbol{a} \tag{7.8}$$

where $\boldsymbol{a}$ is a vector of the unknown parameters and $\mathbf{S}$ is a $p \times p$ covariance matrix. This can be shown by recalling Equation (7.2). For a third degree polynomial, the basis function $b_i(t)$ is a polynomial between the knots, ie. for $t_j < t < t_{j+1}$ it is given by

$$b_i(t) = c_{ij1} + c_{ij2} t + c_{ij3} t^2 + c_{ij4} t^3 \tag{7.9}$$

and the second derivative is

$$b_i''(t) = 2c_{ij3} + 6c_{ij4} t. \tag{7.10}$$

Taking products of terms from (7.10) we have

$$\begin{aligned} b_i''(t) b_l''(t) &= (2c_{ij3} + 6c_{ij4} t)(2c_{lj3} + 6c_{lj4} t) \\ &= 4c_{ij3} c_{lj3} + 12 c_{ij3} c_{lj4} t + 12 c_{ij4} t c_{lj3} + 36 c_{ij4} t c_{lj4} t \\ &= 4c_{ij3} c_{lj3} + (12 c_{ij3} c_{lj4} + 12 c_{ij4} c_{lj3}) t + (36 c_{ij4} c_{lj4}) t^2. \end{aligned} \tag{7.11}$$

To find the elements of the coefficient matrix $\mathbf{S}$ we can first look at one of the intervals

between knots:

$$\begin{aligned}
\int_{t_j}^{t_{j+1}} (\lambda''(t))^2 dt &= \int_{t_j}^{t_{j+1}} \left(\sum_{i=1}^{p} a_i b_i''(t) \right)^2 dt \\
&= \int_{t_j}^{t_{j+1}} \left[\sum_{i=1}^{p} a_i b_i''(t) \right] \left[\sum_{l=1}^{p} a_l b_l''(t) \right] dt \\
&= \int_{t_j}^{t_{j+1}} \left[\sum_{i=1}^{p} \sum_{l=1}^{p} a_i b_i''(t) a_l b_l''(t) \right] dt \\
&= \int_{t_j}^{t_{j+1}} \left[\boldsymbol{a^t} (b_i''(t) b_l''(t)) \boldsymbol{a} \right] dt \\
&= \boldsymbol{a^t} \left[\int_{t_j}^{t_{j+1}} (b_i''(t) b_l''(t)) dt \right] \boldsymbol{a} \\
&= \boldsymbol{a^t} \mathbf{S_j} \boldsymbol{a} \qquad (7.12)
\end{aligned}$$

for i and $l = 1, \ldots, p$ (the number of basis functions) and $j = 1, \ldots, n$ (the number of intervals). $\mathbf{S_j}$ is a matrix of integrals. Then,

$$\mathbf{S} = \sum_{j=1}^{n} \mathbf{S_j}. \qquad (7.13)$$

As mentioned, this gives an improper prior. To make it proper we assume that by age 100 there are no persons still alive. The conditions $\lambda(100) = 0$ and $\lambda'(100) = 0$ would hold; we penalize departures from these assumptions by adding the penalty $\boldsymbol{b_{100}} \boldsymbol{b_{100}^t} + \boldsymbol{b_{100}'} \boldsymbol{b_{100}'^t}$ to $k\mathbf{S}$ to make the improper log prior distribution at Equation (7.8) a proper log prior distribution as follows:

$$f(\mathbf{a}) = -\boldsymbol{a^t} \boldsymbol{\Sigma^{-1}} \boldsymbol{a} \qquad (7.14)$$

where the inverse covariance matrix is

$$\boldsymbol{\Sigma^{-1}} = k\mathbf{S} + \boldsymbol{b_{100}} \boldsymbol{b_{100}^t} + \boldsymbol{b_{100}'} \boldsymbol{b_{100}'^t}. \qquad (7.15)$$

We also assume that $\lambda(t) \geq 0$ for $0 \leq t \leq 100$; we only partially enforce this restriction by truncating the multivariate normal prior so that the condition holds at $t = 0$, 1, 2, ..., 100.

Simulations were used to determine the value of the constant k. The log prior distribution is a truncated multivariate Gaussian distribution with mean vector 0 and covariance matrix $\boldsymbol{\Sigma}^{-1}$. Ten random samples were obtained using the built-in R function `mvrnorm()` for different values of the constant k. Boundary knots were placed at 0 and 100, and the interior knots were placed at the eleven locations of the right-hand boundary of each age group. Figure 7.3 shows ten random samples of $\lambda(t)$ using values of $k = 0.1$, 1, 3 and 10.

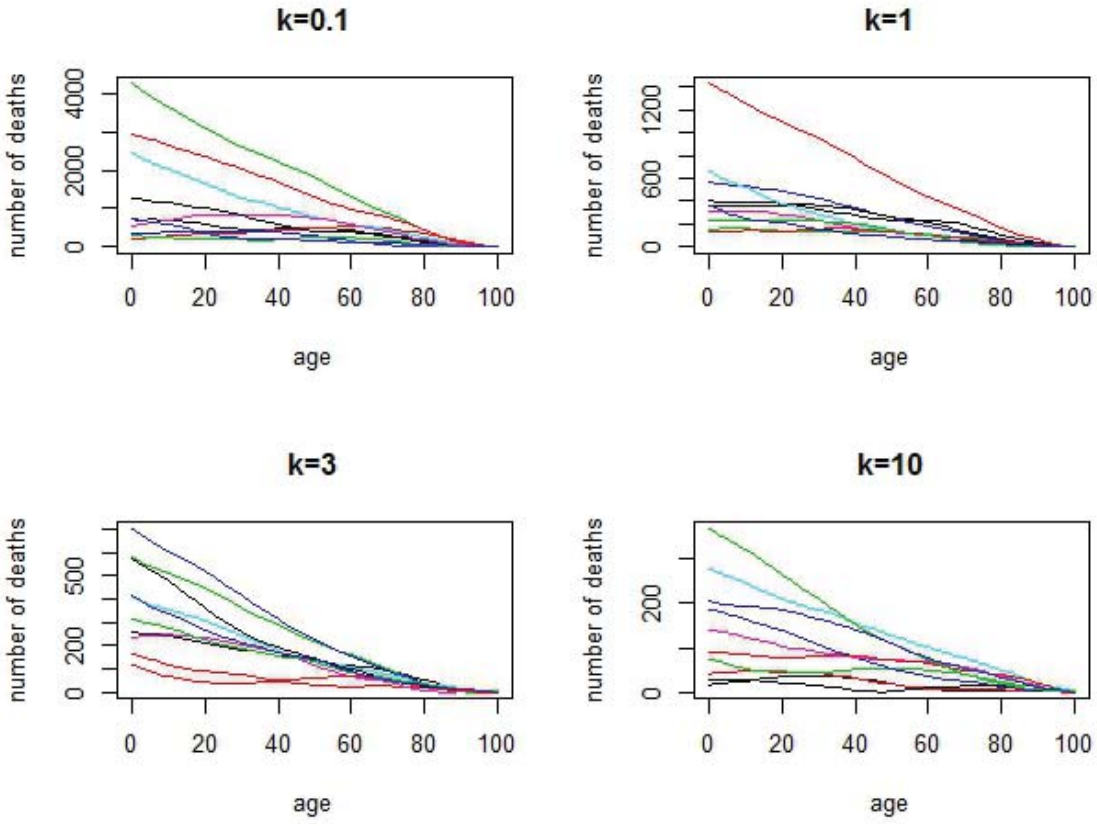

Figure 7.3: Prior samples of $\lambda(t)$ for $k = 0.1$, 1, 3 and 10.

For each value of k the number of deaths decrease with age. As k increases there

does not appear to be any significant change in linearity. The maximum value for the number of deaths is reached when $k = 0.1$. Taking the vertical range into account for each plot, this gives the most curvature for the lowest age group cells when compared to the other values of k. This value of k is used as the initial smoothing parameter.

7.4 Metropolis-Hastings MCMC

A Metropolis-Hastings Markov Chain Monte Carlo (MCMC) algorithm was implemented and used to simulate a sequence of parameters, $\boldsymbol{a}$, of the posterior distribution. The posterior distribution, $f(\boldsymbol{a}|y)$, is called the target distribution. An arbitrary set of values $\boldsymbol{a_0}$ are chosen as a starting point for the parameters. Each iteration of the algorithm follows four steps to update $\boldsymbol{a_l}$ to $\boldsymbol{a_{l+1}}$.:

1. Propose a new value $\boldsymbol{a}^*$ from $h(\boldsymbol{a}^*|y)$ using $\boldsymbol{a}^* = \boldsymbol{a_l} + \mathrm{N_D}(0, \sigma)$, where $\mathrm{N_D}$ is a vector of D independent and identically distributed normals.
2. Calculate $r = \dfrac{f(\boldsymbol{a}^*|y)}{f(\boldsymbol{a_l}|y)}$
3. Sample $u \sim \mathrm{U}[0,1]$.
4. If $u < r$, accept the proposal and set $\boldsymbol{a_{l+1}} = \boldsymbol{a}^*$, otherwise reject the proposal and set $\boldsymbol{a_{l+1}} = \boldsymbol{a_l}$.

The initial set-up for the Metropolis-Hastings MCMC simulation includes a cubic B-spline basis on the interval [0, 100] with knots at the left-hand boundary of each age group except for 0. The dimension of the basis for our model is the number of interior knots plus the degree of the polynomial plus the intercept (D= $11 + 3 + 1 = 15$). The covariance matrix is a 15×15 matrix, and thus, there are 15 unknown parameters. Each $\boldsymbol{a_{li}}$ is the ith component of $\boldsymbol{a_l}$ to form the proposal $\boldsymbol{a}^*$.

Since we assume that the data is Poisson distributed the mean and variance should be equal, however, this is not the case in our data. Multiple observations have been taken across many years. A transformation of the data was made calculating

the mean and variance for each year of the data. If $Y_j \sim \text{Poisson}(\lambda_j)$, then we need to find c such that

$$\text{E}(cY_j) = \lambda_j = \text{Var}(cY_j) \tag{7.16}$$

is true. We tried values of c ranging between 0.01 to 0.08. For our model 0.04 is used for the transformation of the data.

Proposing a random move in step 1 we use a mean of 0 and standard deviation 1. The starting value for each parameter, a_i, is 1. We assume $k = 0.1$ for the constant of the log prior distribution since this value gives the maximum number of deaths for birth given in Figure 7.3. In addition, the condition $\lambda(t) \geq 0$ is implemented.

7.5 Analysis

Using 400,000 iterations with a burn-in of 300,000, the acceptance rate was 19%. Acceptance rates between 15 and 40 percent are ideal (Gelman, Roberts, and Gilks 1996). The sequences for the last 100,000 iterations for each parameter are shown in the trace plots in Figure 7.4. We observe that the sequences improve as the order of the parameters increase. The last plot is ideal since it does not exhibit any pattern or trend. In other words, we want unpredictability which indicates good mixing. We adjusted the standard deviations for each of the parameters are implemented to try and improve the mixing of the lower ordered of parameter values. The standard deviation for each parameter was calulated using the last 100,000 iterations. The standard deviations were (in the order of parameter): 16, 14, 11, 9, 9, 12, 11, 10, 10, 9, 8, 7, 6, 2, and 0.7. Running the simulation using these standard deviations resulted in a low acceptance rate of 0.05%. This is not a surprise since the lower order of parameters converged but did not mix well in the first run of the simulation, and thus, replacing the standard deviations with standard deviations that are known not to mix well gives poor results. The trace plots using different standard deviations for each parameter are shown in Figure 7.5. If the acceptance rate is too low, we shrink

all the jumps proportionally. The model was re-run using the standard deviations (in the order of parameter): 16, 14, 11, 9, 9, 12, 11, 10, 10, 9, 8, 7, 6, 2, and 0.7. divided by 10, the magnitude of the 8th parameter. The was because the 8th parameter gave good results in the trace plot in Figure 7.5. The results of the new trace plots were not improved.

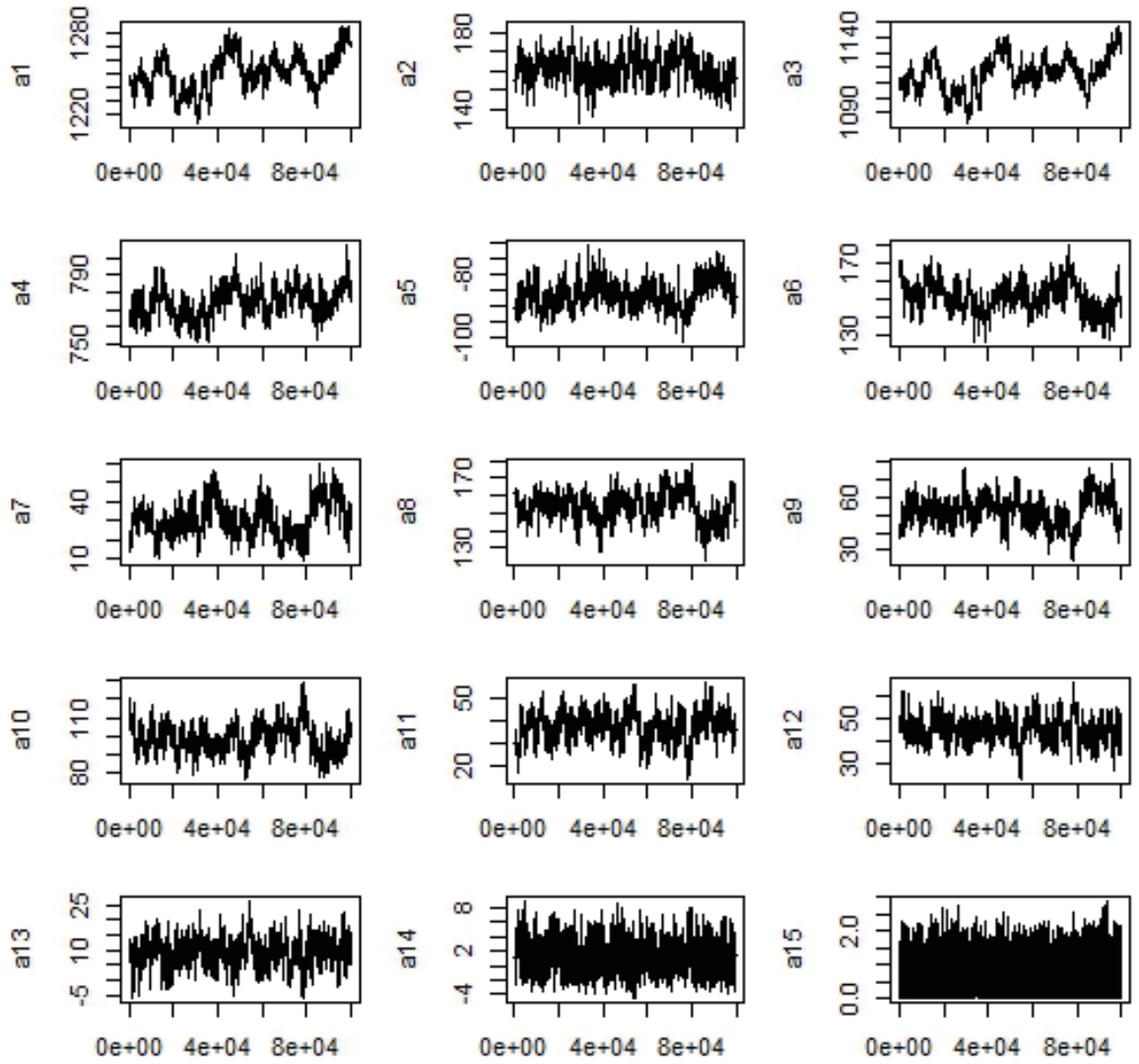

Figure 7.4: Trace plots for the last 100,000 iterations for parameters 1 to 15.

Using the standard deviations from the run of the first simulation, further analysis was explored. The parameters were calculated by computing the mean of the last 100,000 simulations for each sequence of parameter values. The set of means, $\bar{\boldsymbol{a}}$, were

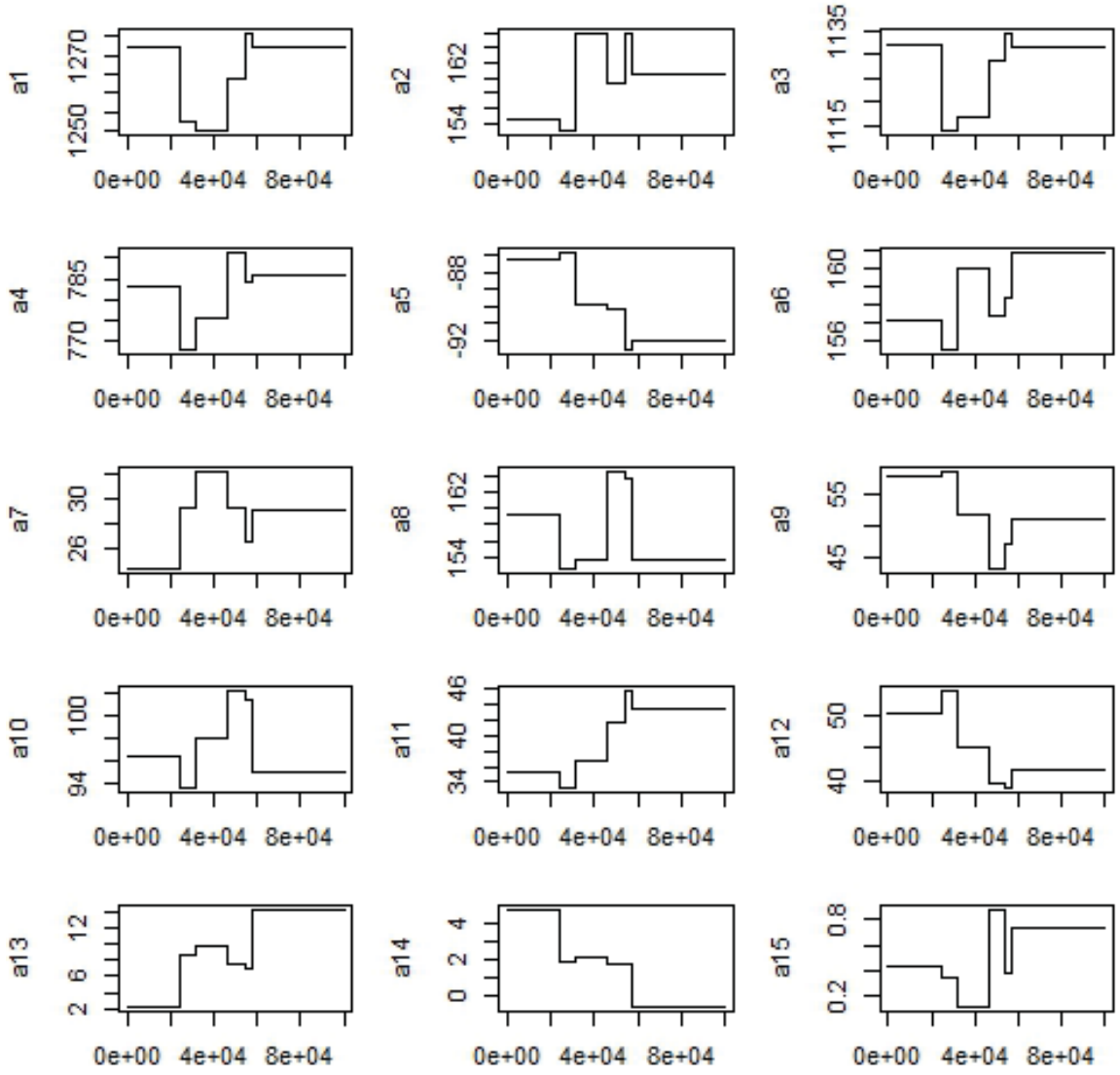

Figure 7.5: Trace plots for the last 100,000 iterations for parameters 1 to 15 using different standard deviations.

used to evaluate the function $\lambda(t)$ and is shown in Figure 7.6. The graph exhibits the features we would expect from a plot showing the number of deaths at each age, ie. starting with a high number of deaths and decreasing as age increases. However, we would expect there to be more curvature between birth to ten years of age because the number of infant deaths was high when the data was collected in the eighteenth century due to disease and illness. Less curvature is required for the higher ages since we would assume less disturbances to effect the number of deaths. Adding more knots

at the lower ages would make very little difference to the fit because curvature is being controlled by k. Figure 7.7 shows the function $\lambda(t)$ and rectangles to represent the data for each age group.

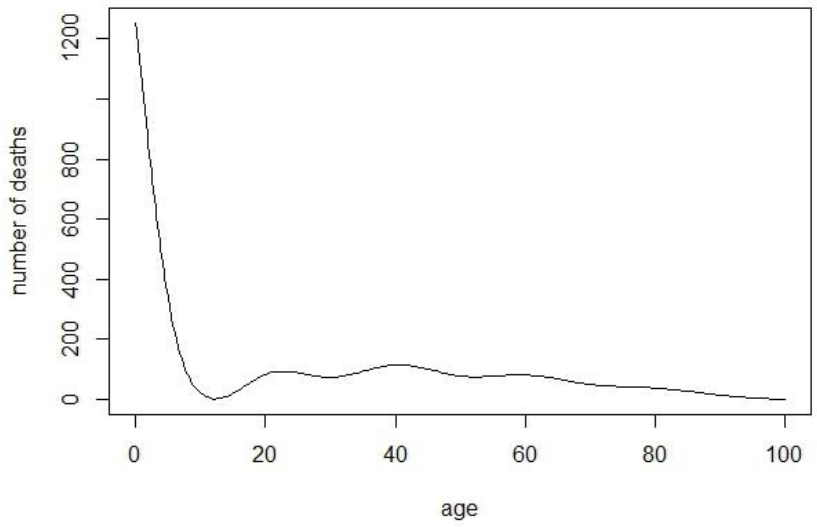

Figure 7.6: $\lambda(t)$: number of deaths per year.

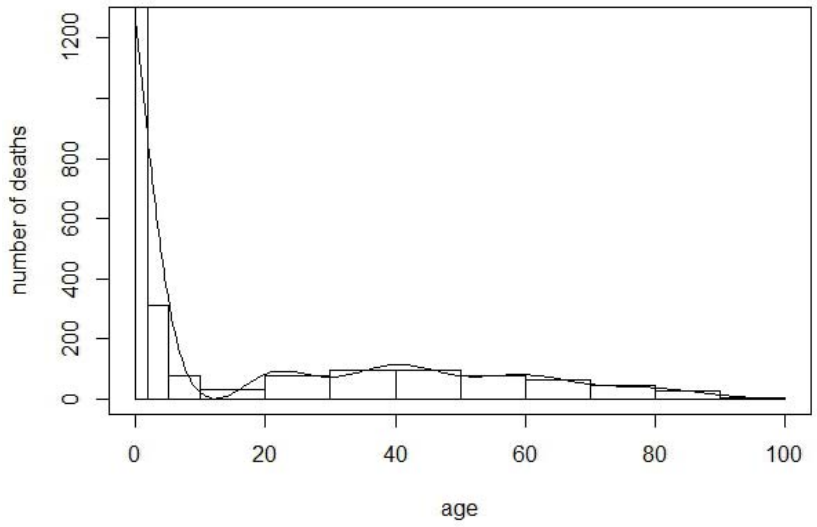

Figure 7.7: $\lambda(t)$ and rectangles representing the area for each age group cell.

The standardized residuals for each age group are shown in Figure 7.8. They were calculated using the following formula:

$$\epsilon_j = \frac{(y_j - \hat{y}_j)}{\sqrt{\hat{y}_j}} \tag{7.17}$$

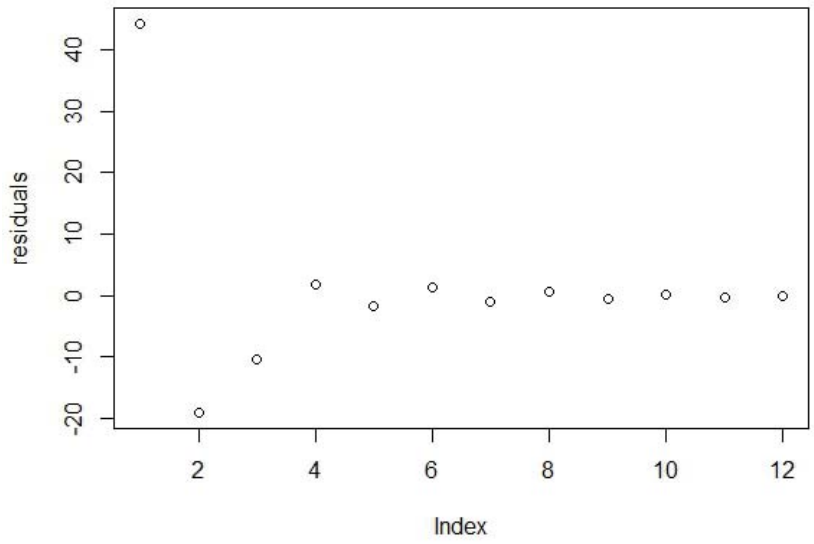

Figure 7.8: Standardized residuals for each age group.

The residuals are large for birth, 2 and 5 years of age. This is caused by having too much penalty at these age intervals.

To improve on the fit of the model, a step function penalty was implemented. As mentioned, less smoothing is acheived when the constant k is small. The value $k = 0.0001$ was used for birth, 2 and 5, and $k = 1$ for the remaining ages. The value $k = 0$ could not be used for the younger age groups since it made the inverse covariance matrix singular, and thus, making the log prior distribution improper.

The acceptance rate using the step function penalty was 25%. The trace plots for the last 100,000 iterations are shown in Figure 7.9 with standard devations for the parameters set to 1. We observe that parameters 1 and 3 show a trend. Different

standard deviations were tested in an attempt to improve the trace plots for parameters 1 and 3. Figure 7.10 show the resulting trace plots using standard devations set to 10, 5 and 5 for the first three parameters respectively. These standard deviations show significant improvement that the model is mixing well. The acceptance rate was 24.6%.

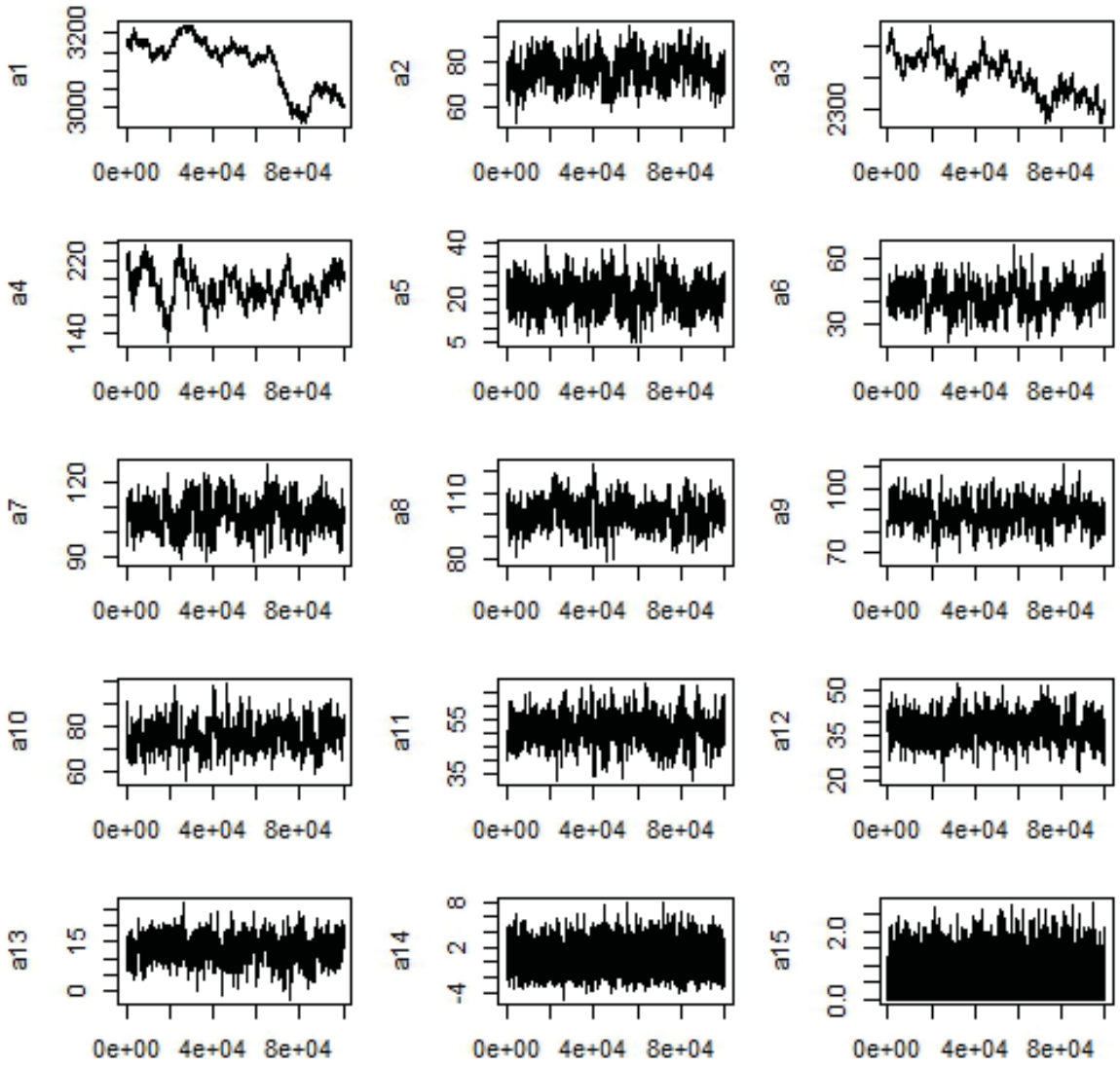

Figure 7.9: Trace plots for the last 100,000 iterations for parameters 1 to 15 using step function penalty.

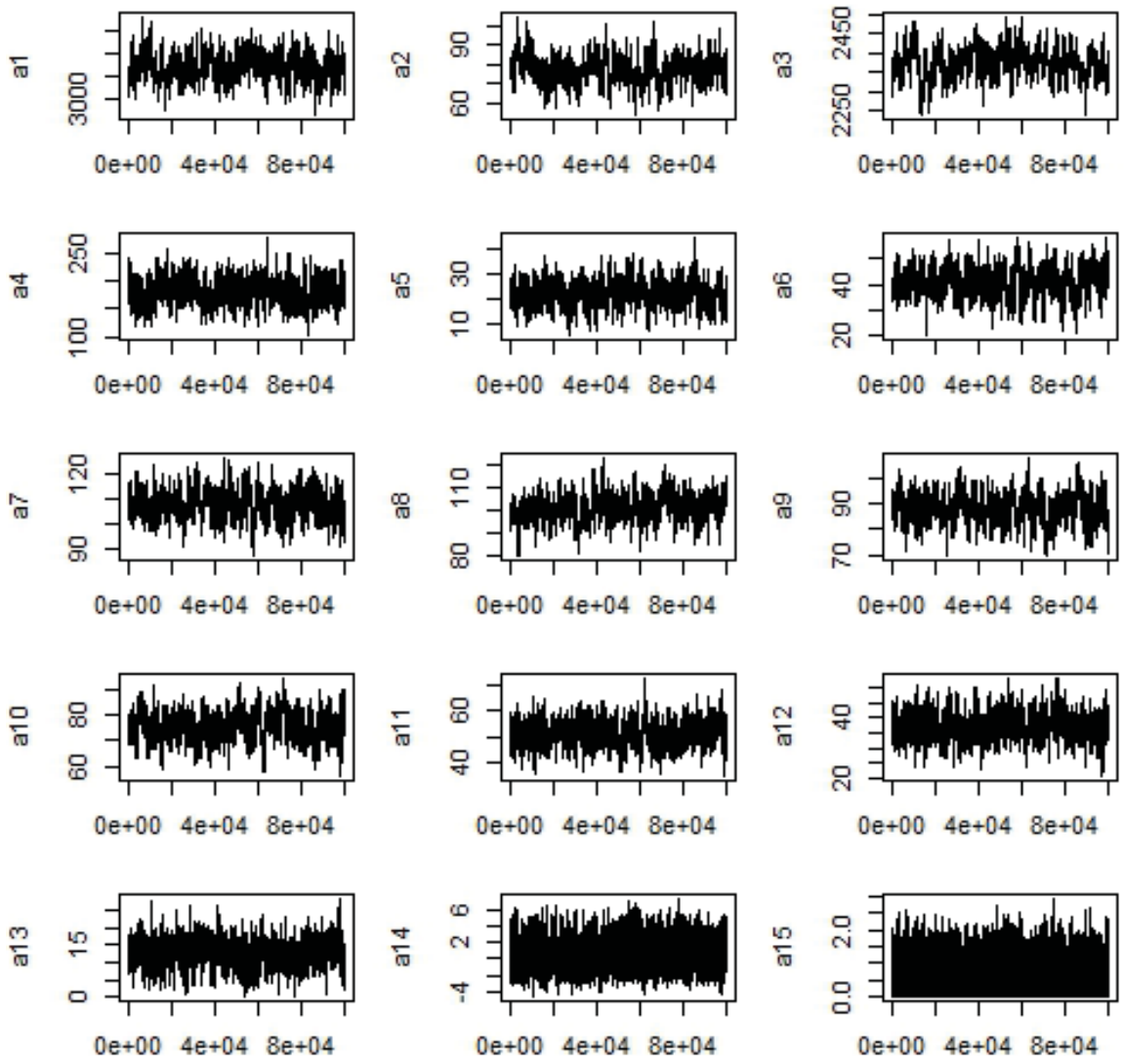

Figure 7.10: Trace plots for the last 100,000 iterations using different standard deviations for parameters 1 to 3.

Figure 7.11 shows the number of deaths (curve) and rectangles representing the data for each age group. Figure 7.12 shows the 95% pointwise credible intervals where the vertical range is limited to 250. There is a 95% probability that $\lambda(t)$ lies between the upper and lower bands. The standardized residuals show an interesting oscillating pattern in Figure 7.13 and Figure 7.14 show the number of deaths (curve) and rectangles representing the data for each age group starting at age 2. We observe that the model has been improved by using the step function for the younger ages.

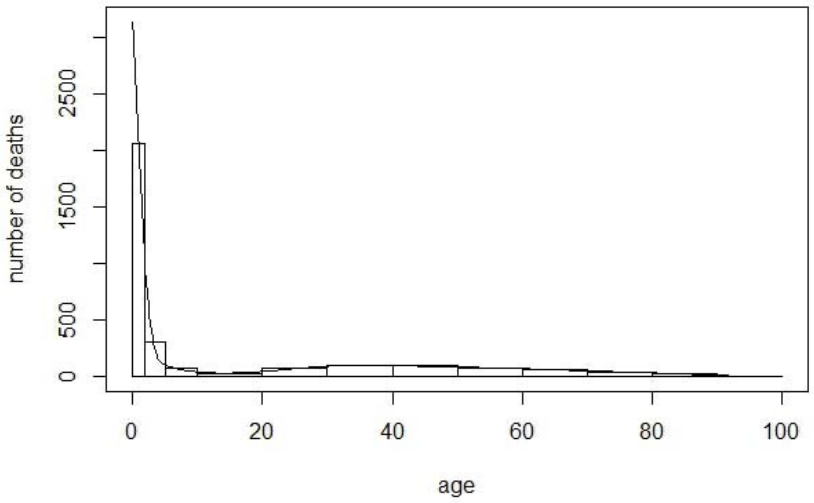

Figure 7.11: $\lambda(t)$ and rectangles representing the area for each age group cell using the step function.

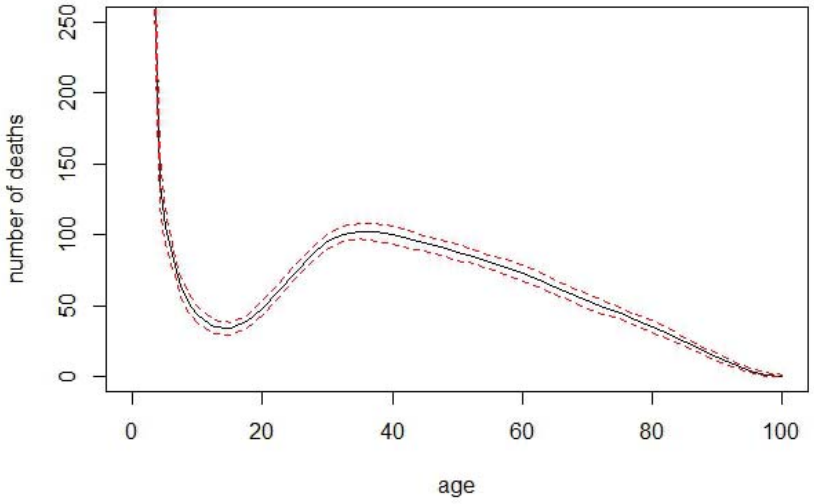

Figure 7.12: $\lambda(t)$ with 95% credible intervals.

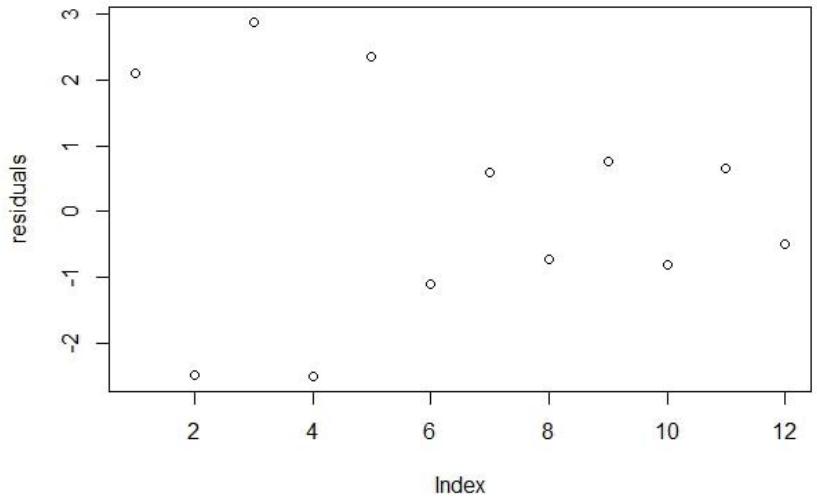

Figure 7.13: Standardized residuals for each age group using step function.

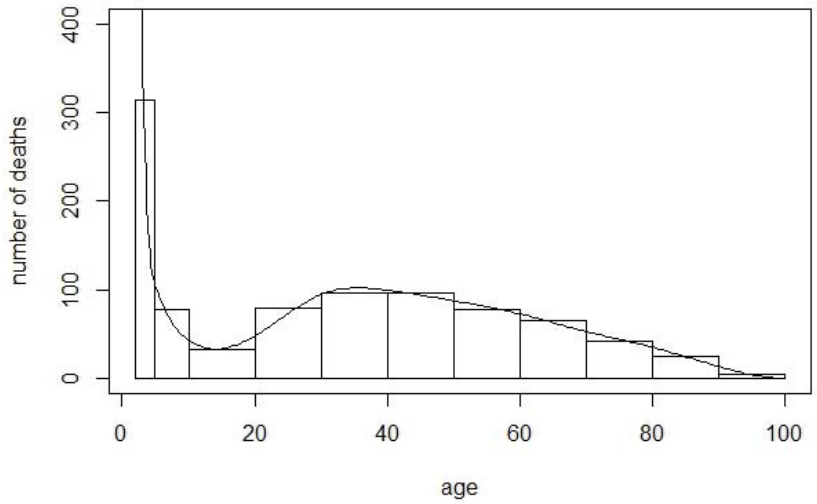

Figure 7.14: $\lambda(t)$ and rectangles representing the area for each age group starting at age 2.

For comparison, the hazard function is shown in Figure 7.15 with 95% credible bands. The hazard function is the instantaneous death rate divided by the survival function:

$$\frac{\lambda(t)}{\int_x^\infty \lambda(t)dt}. \tag{7.18}$$

We observe that starting around age 80 the credible bands begin to widen when compared to the credible bands shown in Figure 7.12.

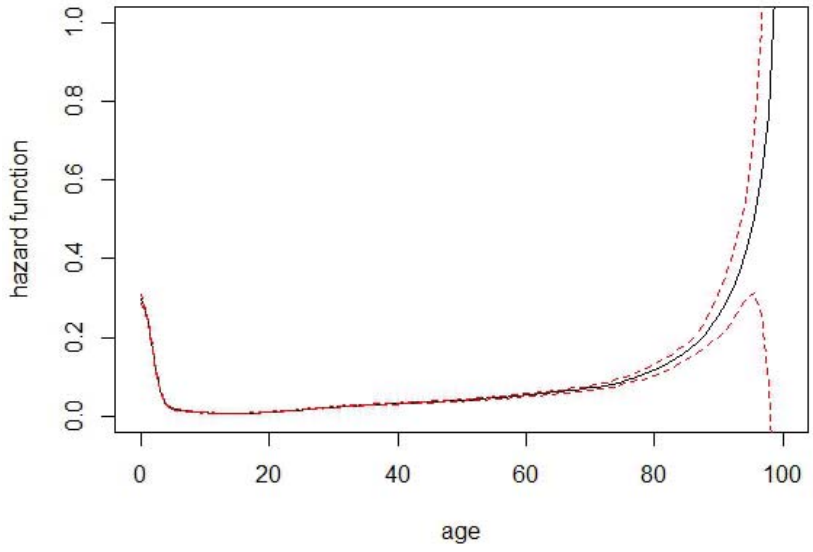

Figure 7.15: Hazard function with 95% credible intervals.

The modern Bayesian smoothing model gives comparable results to that of Smart's life table. Figure 7.16 shows the cumulative number of deaths from the data (solid circles), the Bayesian model (line), and Smart's life table (open circles).

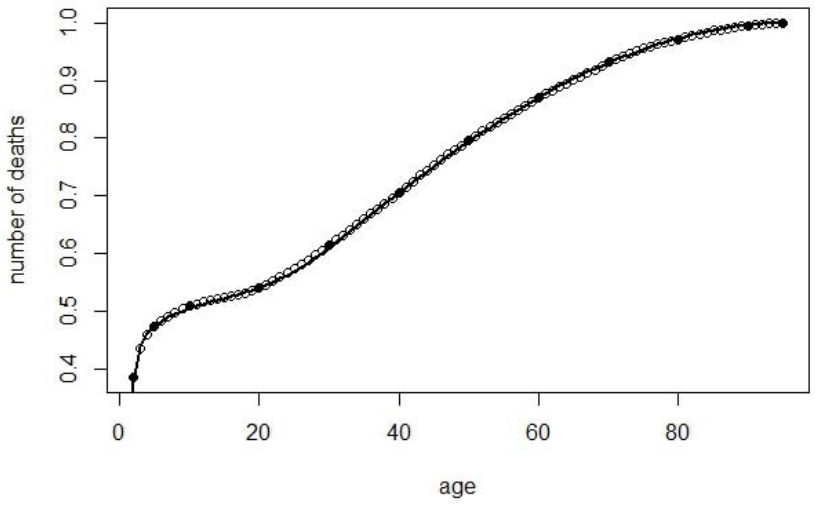

Figure 7.16: Comparison of the Bayesian model (line) and Smart's cumulative distribution (open circles).

7.6 Conclusion

An advantage of using Bayesian smoothing is that the model is flexible. This means the amount of smoothing can be controlled. The location and number of knots can be determined beforehand and assessed using the plots. The trace plots give a graphical way to measure uncertainty and adjust the model as necessary. The acceptance rate is another way to evaluate the performance of the model and to assess if the estimates will be reliable.

This Bayesian model is not an optimal model for smoothing mortality data. However, the purpose of the model was to predict the probability of life using eighteenth-century mortality data and compare the results to the original eighteenth-century analysis. The results show that the Bayesian model works well for this purpose.

Chapter 8

Conclusion

We have looked at some of the significant milestones of data smoothing techniques beginning with Graunt's life table in the seventeenth century. The ingenuity of Halley's life table made it highly influential in various works during the eighteenth century. Elementary smoothing technqiues eventually evolved into complex methods. The large amount of data collected during W.W.I caused an emergence of more advanced methods such as the smoothing method of Sheppard. A rare glimpse into his statistical career and professional relationship with Pearson is provided by the correspondence. Throughout his career, Sheppard retained an interest in the construction of tables based on the normal distribution. The tabulations for his tables would have been computationally intensive and the level of precision demonstrate Sheppard's dedication and skill in practical computation.

Sheppard's smoothing method is not simple to use in practice when compared to other methods such as Spencer's graduation formula. As Sheppard demonstrated, his method has a smaller mean square error and a better fit than Spencer's formula. However, for everyday practical use, simpler methods can provide a good enough approximation to the true values. Sheppard's smoothing method would have been useful for situations that required a high level of precision. This is similiar to his tables based

on the normal distribution having a high number of decimals. For everyday practical use by users of statistics or statisticians, 4 or 5 decimal places is sufficient. However, there are some problems in statistics and science that require rigorous conclusions, and therefore, a higher number of decimals is necessary.

The smoothing methods discussed in this thesis are:

1. Early smoothing techniques: visual or piecewise linear interpolation, averaging (Halley, 1693; Smart, 1738; Simpson, 1742; Price, 1783)
2. Referencing other graduation tables (Smart, 1738; Simpson, 1742; De Moivre, 1725)
3. Graphical methods (Milne, 1815; Sprague, 1886)
4. Parametric models (Gompertz, 1820; Makeham, 1859)
5. Mathematical functions (Graunt, 1662; De Moivre, 1725; Cauchy, 1837)
6. Osculatory interpolation (Sprague, 1886)
7. Difference equation methods (Woolhouse, 1869; Sheppard, 1912–1915)
8. Summation and adjusted averaging (Spencer, 1904)
9. Methods using Mathematical formulae (Farr, 1864; Sheppard, 1912–1915)
10. Local polynomial regression
11. Logistic models
12. Splines, B-splines
13. Bayesian Smoothing

As the collection of detailed population data increased and covered longer periods of time, the methods of design and construction of life tables improved. Early life table compilers faced the challenge of determining estimates using sparse or incomplete data. It is the same issue today when working with data from countries that do not reliably collect and record vital statistics. As discussed, advanced techniques such as Bayesian smoothing can be implemented to address the issue.

References

Book Sources

Bacaër, Nicholas (2011). *A Short History of Mathematical Population Dynamics*. London: Springer-Verlag London Limited.

Bellhouse, David R. (2011b). *Abraham De Moivre: Setting the Stage for Classical Probability and its Applications*. Florida: CRC Press Taylor & Francis Group.

Brown, William (1921). *The Essentials of Mental Measurement*. London: Cambridge.

De Moivre, A. (1725). *Annuities upon Lives, or; The valuation of Annuities upon any Number of Lives, as also, of Reversions to which is added, an Appendix Concerning the Expectations of Life, and Probabilities of Survivorship*. London: W.P.

Elderton, William P. and Norman L. Johnson (1969). *Systems of Frequency Curves*. London: Cambridge.

Farr, William (1864). *The English Life Table*. London: Longman Green and Roberts.

Fisher Box, Joan (1978). *R.A. Fisher: The Life of a Scientist*. New York: John Wiley & Sons.

Hald, Anders (1990). *A History of Probability and Statistics and Their Applications before 1750*. New York: John Wiley & Sons.

Lewin, Christopher and Margaret De Valois (2003). *The History of Mathematical Tables: From Sumer to Spreadsheets*. Ed. by R. Flood M Cambell-Kelly M. Croarken and E. Robson. Chapter title: "History of actuary tables". Oxford: Oxford University Press.

Lindley, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*. London: Cambridge University Press.

Macaulay, F.R. (1931). *The Smoothing of Time Series*. New York: National Bureau of Economic Research.

MacKenzie, Donald A. (1981). *Statistics in Britain 1865-1930: The Social Construction of Scientific Knowledge*. Edinburgh: Edinburgh University Press.

Milne, Joshua (1815). *A Treatise on the Valuation on Annuitities and Assurances on Lives and Survivorships, Volumes 1 and 2*. London: Longman, Hurst, Rees, Orme, and Brown.

Pearson, E.S. and H.O. Hartley (1970). *Biometrika Tables for Statisticians*. London: Cambridge University Press.

Pearson, K. (1914a). *Tables for Statisticians and Biometricians Volume 1*. Ed. by K. Pearson. London: Cambridge University Press.

Pearson, K. (1914b). *Tables for Statisticians and Biometricians Volume 2*. Ed. by K. Pearson. London: Cambridge University Press.

Pearson, K. (1924). *Tables for Statisticians and Biometricians*. Ed. by K. Pearson. London: Cambridge University Press.

Porter, Theodore M. (2004). *Karl Pearson*. New Jersey: Princeton University Press.

Price, Richard (1783). *Observations reversionary payments; schemes for providing annuities for widows, and for persons in old age; on the method of calculating the values of assurances on lives, 4th edition*. Vol. I. London: T. Cadell.

Registrar-General (1848). *8th Report of the Registrar General: Births, Deaths and Marriages*. Wiliiam Cloives and Sons.

Registrar-General (1853). *12th Report of the Registrar General: Births, Deaths and Marriages*. London: Wiliiam Cloives and Sons.

Rhodes, E.C. (1921). *Tracts for Computers, No. VI Smoothing*. Ed. by K. Pearson. London: Cambridge University Press.

Savage, L.J. (1954). *The Foundations of Statistics*. New York: Wiley.

Sheppard, W.F. (1939). *The Probability Integral, British Association of Mathematical Tables*. London: Cambridge University Press.

Simpson, Thomas (1742). *The doctrine of annuities and reversions, deduced from general and evident principles*. London: Printed for J. Norse.

Smart, John (1726). *Tables for Interest, Discount, Annuities, & c.* London: Darby and Brown.

Wahba, Grace (1990). *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

Articles

Aitken, A.C. (1938). "A Note on Sheppard's Contribution to Mathematics and Mathematical Statistics". In: *Annals of Eugenics* 8, pp. 9–11.

Anderson, Oskar (1927). "On the Logic of the Decomposition of Statistical Series into Separate Components". In: *Journal of the Royal Statistical Society* 90.3, pp. 548–569.

Bellhouse, David R. (1998). "London Plague Statistics in 1665". In: *Journal of Official Statistics* 14.2, pp. 207–234.

Bellhouse, David R. (2011a). "A new look at Halley's life table". In: *Journal of the Royal Statistical Society: Series A* 174.3, pp. 823–832.

BMJ (1902). "On the Logic of the Decomposition of Statistical Series into Separate Components". In: *British Medical Journal* 1.2142, p. 161.

Carlin, B.P. (1992). "A simple Monte-Carlo approach to Bayesian graduation". In: *Transactions of Society of Actuaries* 44, pp. 55–76.

Cauchy, A.L. (1837). "Sur l'interpolation". In: *Journal de Mathématiques Pures et Appliquées. Paris* 2, pp. 193–205.

Congdon, Peter (2009). "Life expectancies for small areas: a Bayesian random effects methodology". In: *International Statistical Review* 77.2, pp. 222–240.

Cornfield, Jerome and Katherine Detre (1977). "Bayesian life table analysis". In: *Journal of the Royal Statistical society Series B* 39.1, pp. 86–94.

Croarken, M. and Martin Campbell-Kelly (2000). "Beautiful numbers: the rise and decline of the British Association Mathematical Tables Committee, 1871-1965". In: *Annals of Eugenics* 22.4, pp. 44–61.

De Finetti, B. (1961). "The Bayesian approach to the rejection of outliers". In: *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.* 1, pp. 199–210.

De Finetti, B. (1974). "Bayesianism: Its unifying role for both the foundations and applications of statistics". In: *International Statistical Review* 42.2, pp. 117–130.

Dellaportas, Petros, Adrian F.M. Smith, and Photis Stavropoulos (2001). "Bayesian analysis of mortality data". In: *Journal of the Royal Statistical Society Series A* 164.2, pp. 275–291.

Galton, F. (1907). "Grades and Deviates: (Including a table of normal deviates corresponding to each millesimal grade in the length of an array, and a figure)". In: *Biometrika* 5.4, pp. 400–406.

Gelman, O., G.O. Roberts, and W.R. Gilks (1996). "Efficient Metropolis Jumpting Rules". In: *Bayesian Statistics* 5, pp. 599–601.

Geman, S. and D. Geman (1984). "Stochastic relaxation, gibbs distributions, and the Bayesian resoration of images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, pp. 721–741.

Gompertz, Benjamin (1820). "A sketch of an analysis and notation applicable to the estimation of the value of life contigencies". In: *Philosophical Transactions* 110, pp. 214–294.

Gompertz, Benjamin (1825). "On the nature of the function expressive of the Law of Human Mortality, and on a new mode of determining the value of life contingencies". In: *Philosophical Transactions* 115, pp. 513–583.

Gompertz, Benjamin (1861). "Supplement to two papers published in the Philosophical Transactions (1820 and 1825) on the science connected with human mortality". In: *Proceedings of the Royal Society of London* 11, pp. 390–392.

Haario, H., E. Saksman, and J. Tamminen (2001). "An adaptive Metropolis algorithm". In: *Bernoulli* 7, pp. 223–242.

Halley, Edmond (1693). "An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the City of Breslaw; with an attempt to ascertain the price of annuities upon lives". In: *Philosophical Transactions* 17, pp. 596–610.

Hartley, H.O. (1940). "The Probability Integral by W.F. Sheppard Review". In: *Journal of the Royal Statistical Society* 103, pp. 94–96.

Hastings, W.K. (1970). "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57, pp. 97–109.

Henderson, Robert (1916). "Note on graduation by adjusted average". In: *Actuarial Society America* 17, pp. 43–48.

Henderson, Robert (1924). "A new method of graduation". In: *Actuarial Society America* 25, pp. 29–39.

Hickman, J.C. and R.B. Miller (1977). "Notes on Bayesian graduation". In: *Transactions of Society of Actuaries* 29, pp. 1–21.

Jeffreys, Harold (1946). "An invariant form for the prior probability in estimation problems". In: *Proceedings of the Royal Society of London Series A* 186, pp. 453–461.

Kimeldorf, G.S. and Donald A. Jones (1967). "Bayesian graduation". In: *Transactions of Society of Actuaries* 19.54, pp. 66–127.

Kimeldorf, G.S. and Grace Wahba (1970). "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines". In: *The Annals of Mathematical Statistics* 41.2, pp. 495–502.

King, George (1883). "On the method used by Milne in the construction of the Carlisle Table of Mortality". In: *Journal of the Institute of Actuaries* 24, pp. 186–211.

Laplace, Pierre Simon (1774). "Mémoire sur la probabilité des causes par les événements". In: *Mém. de l'Académie royale des sciences presentés par divers savans* 6. Translated into English with an introduction by S.M. Stigler in *Statistical Science*, 1986, 1, 359-378., pp. 621–656.

Laplace, Pierre Simon (1781). "Mémoire sur la probabilité". In: *Mém. de l'Académie royale des sciences de Paris* 9, pp. 383–485.

Luoma, Arto, Anne Puustelli, and Lasse Koskinen (2012). "A Bayesian smoothing spline method for mortality modelling". In: *Annals of Actuarial Science* 6, pp. 284–306.

Makeham, William M. (1859). "On the law of mortality and the construction of annuity tables". In: *Journal of the Institute of Actuaries* 8, pp. 301–310.

Medical, The Edinburgh and Surgical Journal (1817). "Statement on the sizes of men in different counties of Scotland". In: *The Edinburgh Medical and Surgical Journal* 13, pp. 260–262.

Metropolis, N. et al. (1953). "Equations of state calculations by fast computing machines". In: *Journal of Chemical Physics* 21, pp. 1087–1092.

Newman, F.W. (1883). "Table of the descending exponential function to twelve or fourteen places of decimals". In: *Transactions of the Cambridge Philosophical Society* 13, pp. 145–241.

Pearson, K. (1895). "Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material". In: *Philosophical Transactions A* 186, pp. 343–414.

Pearson, K. (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonable supposed to have arisen from random sampling". In: *Philosophical Magazine Series 5* 50, pp. 157–175.

Pearson, K. (1902). "On the systematic fitting of curves to observations and measurements". In: *Biometrika* 1, pp. 265–303.

Pearson, K. (1903). "On the probable errors of frequency constants". In: *Biometrika* 2.3, pp. 273–281.

Pearson, K. and L.N.G. Filon (1898). "Mathematical contributions to the theory of evolution. IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation". In: *Philosophical Transactions A* 191, pp. 229–311.

Pearson, K. and Alice Lee (1908). "On the generalised probable error in multiple normal correlation". In: *Biometrika* 6.1, pp. 59–68.

Price, Richard and Thomas Bayes (1763). "An Essay towards solving a problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F.R.S. Communicated by Mr. Price, in a Letter to John Canton, A.M.F.R.S." In: *Philosophical Transactions* 53, pp. 370–418.

Robert, Christian and George Casella (2011). "A short history of Markov chain Monte Carlo: subjective recollections from incomplete data". In: *Statistical Science* 0, pp. 1–14.

Sheppard, N.F. (1938). "W.F. Sheppard: Personal History". In: *Annals of Eugenics* 8, pp. 1–9.

Sheppard, W.F. (1889). "On some expressions of a function of a single variable in terms of Bessel's functions". In: *Quarterly Journal of Pure and Applied Mathematics* 23, pp. 223–260.

Sheppard, W.F. (1897a). "On the calculation of the average square, cube & c., of a large number of magnitudes". In: *Journal of the Royal Statistical Society* 60, pp. 698–703.

Sheppard, W.F. (1897b). "On the geometrical treatment of the 'normal curve' of statistics". In: *Proceedings of the Royal Society of London* 62, pp. 170–173.

Sheppard, W.F. (1898). "On the calculation of the most probable values of frequency constants, for data arranged according to equidistant divisions of a scale". In: *Proceedings of the London Mathematical Society* 29, pp. 353–380.

Sheppard, W.F. (1899a). "A method for extending the accuracy of certain mathematical tables". In: *Proceedings of the London Mathematical Society* 31, pp. 423–448.

Sheppard, W.F. (1899b). "Central-Difference Formulae". In: *Proceedings of the London Mathematical Society* 31, pp. 479–482.

Sheppard, W.F. (1899c). "On the application of the theory of error to cases of normal distribution and normal correlation". In: *Philosophical Transactions A* 192, pp. 101–167.

Sheppard, W.F. (1900). "Some quadrature-formulae". In: *Proceedings of the London Mathematical Society* 32, pp. 258–277.

Sheppard, W.F. (1903). "New tables of the probability integral". In: *Biometrika* 2, pp. 174–190.

Sheppard, W.F. (1912). "Reduction of errors by means of negligible differences". In: *Proceedings of the Vth International Congress of Mathematics, Cambridge* 2, pp. 348–384.

Sheppard, W.F. (1914a). "Fitting of polynomial by method of least squares". In: *Proceedings of the London Mathematical Society* 13, pp. 97–108.

Sheppard, W.F. (1914b). "Graduation by reduction of mean square error Part I". In: *Journal of the Institute of Actuaries* 48, pp. 171–185.

Sheppard, W.F. (1914c). "Graduation by reduction of mean square error Part II". In: *Journal of the Institute of Actuaries* 48, pp. 390–412.

Sheppard, W.F. (1915). "Graduation by reduction of mean square error Part III". In: *Journal of the Institute of Actuaries* 49, pp. 148–157.

Sheppard, W.F. (1929). "The fit of a formula for discrepant observations". In: *Philosophical Transactions A* 228, pp. 115–150.

Sherriff, Catherine W.M. (1920). "On a Class of Graduation Formulae". In: *Proceedings of the Royal Society of Edinburgh* 40, pp. 112–128.

Smart, John (1738a). "A letter from Mr. Smart to George Heathcote Esq., re inclosing tables extracted from ye bills of mortality for the last ten years, and from them shewing the probabilities of life, in order to estimate annuities &c". In: *Journal Book of Scientific Meetings of the Royal Society*.

Spencer, John (1904). "On the graduation of the rates of sickness and mortality presented by the experience of the Manchester Unity of Oddfellows during the period 1893–97". In: *Journal of the Institute of Actuaries* 38, pp. 334–343.

Sprague, T. (1886). "The graphic method of adjusting mortality tables. A description of its objects, and its advantages as compared with other methods, and an application of it to obtain a graduated mortality table from Mr. A.J. Finlaison's observations on the mortality of the female government annuitants, 4 years and upwards after purchase". In: *Journal of the Institute of Actuaries* 26, pp. 77–120.

Stigler, Stephen (2008). "Karl Pearson's theoretical errors and the advances they inspired". In: *Statistical Science* 23, pp. 261–274.

Student (1908). "The probable error of a mean". In: *Biometrika* 6.1, pp. 1–25.

Sutton, William (1883). "On the method used by Dr. Price in the construction of the Northampton Mortality Table". In: *Journal of the Institute of Actuaries* 18, pp. 107–122.

Vernon Boys, Charles (1944). "Obituary Notices of Fellows of the Royal Society". In: *The Royal Society* 4.13, pp. 771–788.

Wahba, Grace (1978). "Improper priors, spline smoothing and the problem of guarding against model errors in regression". In: *Journal of the Royal Statistical Society Series B* 40.3, pp. 364–372.

Whittaker, E.T. (1923). "On a new method of graduation". In: *Proceedings of the Edinburgh Mathematical Society* 41, pp. 63–75.

Woolhouse, W.S.B. (1869). "Explanation of a new method of adjusting mortality tables; with some observations upon Mr. Makeham's modification of Gompertz's Theory". In: *Journal of the Institute of Actuaries* 15, pp. 389–410.

Other Sources

Fisher, R. (1936). *Fisher Correspondence*. Correspondence from Fisher to W.F. Sheppard. URL: `http://hdl.handle.net/2440/67994` (visited on 09/04/2014).

Fisher, R. (1937). *Fisher Correspondence*. Correspondence from Fisher to N.F. Sheppard. URL: `http://hdl.handle.net/2440/67992` (visited on 09/04/2014).

Graunt, John (1662). "National and political observations made upon the bills of mortality". In: Bills of Mortality, London, pp. 23–127.

Pearson, Karl (1896–1926). "Sheppard, William Fleetwood". In: *Pearson Papers*. Correspondence from W.F. Sheppard to Pearson. University College London Archives, Special Collections, Pearson/11/1/18/77.

Smart, John (1738b). "A Table Showing the Probabilities of Life". In: *The Bills of Mortality for the City of London*. Guildhall Library, London, England.

Statistics-Canada (2015). *Methods for Constructing Life Tables for Canada, Provinces and Territories*. URL: `http://www.statcan.gc.ca/pub/84-538-x/84-538-x2013001-eng.htm` (visited on 12/02/2015).

Appendix A

Smart's Life Table

John Smart's table (1738) showing the probabilities of life by observations made from the Bills of Mortality for the City of London from 1728–1737.

Age	*Live*	*Deaths*	*Age*	*Live*	*Deaths*
Born	1,000	0	10	490	5
1	710	290	11	486	4
2	614	96	12	482	4
3	564	50	13	479	3
4	539	25	14	477	2
5	526	13	15	475	2
6	516	10	16	473	2
7	508	8	17	471	2
8	501	7	18	468	3
9	495	6	19	464	4
Age	*Live*	*Deaths*	*Age*	*Live*	*Deaths*
20	459	5	30	385	9
21	453	6	31	376	9
22	447	6	32	367	9
23	440	7	33	358	9
24	433	7	34	349	9
25	426	7	35	340	9
26	418	8	36	331	9
27	410	8	37	322	9
28	402	8	38	313	9
29	394	8	39	304	9

Age	*Live*	*Deaths*	*Age*	*Live*	*Deaths*
40	294	10	50	204	8
41	284	10	51	196	8
42	274	10	52	188	8
43	264	10	53	180	8
44	255	9	54	172	8
45	246	9	55	165	7
46	237	9	56	158	7
47	228	9	57	151	7
48	220	8	58	144	7
49	212	8	59	137	7
Age	*Live*	*Deaths*	*Age*	*Live*	*Deaths*
60	130	7	70	69	6
61	123	7	71	64	5
62	117	6	72	59	5
63	111	6	73	54	5
64	105	6	74	49	5
65	99	6	75	45	4
66	93	6	76	41	4
67	87	6	77	38	3
68	81	6	78	35	3
69	75	6	79	32	3
Age	*Live*	*Deaths*	*Age*	*Live*	*Deaths*
80	29	3	90	5	1
81	26	3	91	4	1
82	23	3	92	3	1
83	20	3	93	2	1
84	17	3	94	1	1
85	14	3	95	0	1
86	12	2			
87	10	2			
88	8	2			
89	6	2			

Appendix B

Correspondence from W.F. Sheppard to K. Pearson

I have transcribed and included footnotes for 23 letters from W.F. Sheppard to K. Pearson archived at University College London, London (UCL Archives Special Collections PEARSON/11/1/18/77). The letters are presented in chronological order.

Letter 1

2 Temple Gardens, E.C.

3 June '96

Dear Sir,

I believe Mr. Galton spoke to you a short time ago with regard to our unfinished paper of mine on the "normal curve" in relation to statistics. He thought that you might be willing to look through it when finished, in order to see whether it would be suitable for the Royal Society, and whether, if submitted to them, there would be any chance of it being published in the Phil. Trans.[1]

I had hoped to get on with the paper during the spring, but circumstances have prevented my doing so, and there does not seem much probability of my completing it before the end of the year. If, however, you would be kind enough to look at the paper, I do not see why I should not send you what I have already done, so that you

[1]Paper was first published in 1897 "On the geometrical treatment of the 'normal curve' of statistics". in *Proceedings of the Royal Society of London* 62, pp. 170–173, and revised and republished in 1899 "On the application of the theory of error to cases of normal distribution and normal correlation" in *Philosophical Transactions* A 192, pp. 101–167.

might read it at your leisure, and I could at the same time give a sketch of what I propose to add in order to complete the paper.

It deals almost entirely with the normal curve, and with correlation between normal distributions, non-normal distributions being only considered for the purpose of analysing them into component normal distributions. As regards results, there is a good deal of new matter in the part relating to correlation; but the work is entirely theoretical, & the greater part of the paper consists in the application of geometrical methods so as to obtain results which are already known. On this account I have been doubtful whether the paper was really suitable for the Phil. Trans.

The portion already finished occupies a good deal of space, but I have purposely treated the subject thoroughly. As I have wished to make it intelligible to others besides the few who have so far worked at it. You would read through it very rapidly, and I think you would find a good deal of the geometrical work interesting, though I admit that some is rather tedious. There is no use of diffn. or int., except by geometrical methods, though there is a certain amount of analytical work to be added.

I could send you the paper by post, or could call on you some time & discuss the matter with you, if you prefer that. Almost any time would suit me, as I have few definite engagements just at present.

Yours faithfully,

W.F. Sheppard

I may add that I should be much gratified if any of my work would be of use to you in your own investigations.

Letter 2

2 Temple Gardens, E.C.

16 June '96

Dear Mr. Pearson,

Thanks for the Czuber,[2] which—when rid of the "theory of error" jargon—ought to be interesting. I will return it to you by Aug. 1st. It will be a good training for any summer holiday in Germany.

I should be very pleased of any work that you can put in any way, either private coaching or classwork. My financial condition precludes the single-minded devotion to marginal annotations, which is necessary for success at the bar; and even if this

[2]Emanuel Czuber was an Austrian mathematician. Pearson had probably sent Sheppard a book. The mostly likely candidate is one published in 1891, *Theorie der Beobachtungsfehler* Teubner, Leipzig. English translation, as *The theory of errors of observation.*

were not so, I think I should be wanting to spread myself over other things. So I hope ultimately to get some permanent post in London, probably something involving administrative educational work, and in the mean time I am anxious to get as much teaching work as I can.

Yours very truly,

W.F. Sheppard

I am ready for pupils all through the summer, except definitely for two or three weeks in August.

Letter 3

2 Temple Gardens, E.C.

19 Oct. 1896

Dear Mr. Pearson,

I send you this M.S., on an isolated point, as I think you may like to see it before it is published. When I wrote the first draft of it I did not know that you had gone into the subject at all, as I had not read your "Skew Variation" essay thoroughly.[3] I have since introduced a reference to this, and taken for illustration one of your tables. You will see that my result is very different from yours, and mine seems the more correct. Where the flaw in your reasoning comes in is that you take the ordinate y_r, as proportional to the number n_r. This is correct for a first approx., but a 1st approx. corresponds to a figure of frequency composed of rectangles. A polygon is equivalent to a 2nd approx., & for a 2nd approx., the ordinate would be $n_r + \frac{1}{2}[n_r - \frac{1}{2}(n_{r-1} + n_{r+1})]$. Hence by taking it ($= n_r$) you make it too small where the curve is concave to the base, i.e. (usually) at the centre, & too great where the curve is convex to the base, i.e. at the extremities, so that the value obtained for the S.D. be too great. (Instead of the central ordinate of the compartment I find it more convenient to deal with the bounding ordinates $\frac{1}{2}(n_{r-1} + n_r)$ and $\frac{1}{2}(n_r + n_{r+1})$, but the result would be the same to this order of approx.) Its to a 3rd approx., introducing Δ^2, but his happens to make no alteration in the value of the average, though it introduces a small term into the S.D.

I have put this M.S. in as untechnical language as possible, as I thought it might be suitable for the statistical society. I am putting the mathematical part into a separate paper, which will give the corresponding formula for the nth moment accurately (your M'_n). This might be suitable for the Phil. Mag. or the Cambridge

[3]Sheppard is referring to Pearson's 1885 paper, "Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material" in *Philosophical Transactions* A 186, pp. 343–414.

Phil. Soc.[4] Perhaps you could advise me about this.

Yours sincerely,

W.F. Sheppard. (P.T.O.)

P.S. Can you tell me what is the recognised scale of fees for mathematical coachings? I have been in the habit of charging 10/6 an hour, but it has been suggested to me that this is too high for elementary work, e.g. Cambridge Little–So, & that it should only be about 7/ an hour. You ought to be in a position to know-I have enquired of one or two people, but they have not been able to tell me. W.F.S.

Letter 4

2 Temple Gardens,

20 Oct. '96

Dear Mr. Pearson,

I find it rather hard to give an answer with regard to the lecturing in Astronomy. The principal reason is that I am not certain what I may be doing after Christmas: it is no use counting one's chickens etc., but it is not fair to undertake work & find oneself compelled to give it up very soon. Partly for this reason, and partly because I want to get on as far as possible before Christmas with what I can prepare for publication, I was anxious to keep free just at present from any distant engagements. A secondary reason—which however could be overcome—is that Astronomy is just the one subject in which the smattering acquired at Cambridge has failed to interest me; and a smart student would discover my ignorance at once. But if you find yourself in a difficulty with regard to getting a lecturer, I shall be glad to do what I can; though, as I have explained, I feel inclined to hope I amy not be called upon.

I am much obliged for your explanation of your method of "fitting" the curve. I must confess I had misunderstood it. I had not realised that you sought a curve which was not the curve of frequency but was related to it in a particular way: and I will modify the note in my MS accordingly. But I don't yet see what your curve is in the case of (e.g.) a normal distribution; apparently the ordinate y_r of the curve at every point x_r is to be proportional to the area of the curve of frequency between $x_r - \frac{1}{2}c$ and $x_r + \frac{1}{2}c$; but there is no finality in this, as the shape of the new curve will depend on the value of c. Apparently we look at the thing in different ways. I do not try to find a frequency curve of which the numbers given could be successive areas: I try to find the frequency curve which would result if the causes or whatever

[4]Sheppard's manuscript was published in 1898 "On the calculation of the most probable values of frequency constants, for data arranged according to equidistant divisions of a scale" in *Proceedings of the London Mathematical Society* 29, pp. 353–380. He refers to Pearson's method of moments in the Appendix title "Moments of a Polygon" pp. 378–380

they are—which regulate the particular magnitude in the individuals measured acted in the same way on an infinite number of individuals, the hypothesis being that the particular individuals are a chance selection from this infinite number. My language is neither philosophical nor clear (this is not a pleonasm), but the hour is late for Proc. R.S. And then the geometrical discussion of "normal" distributions *ab initio*, which would have to be relegated to some Transactions. I find that while working out formulae on scribbling—paper takes some time, bringing them into intelligible language takes a good deal more, and numerical illustrations take still more, so that my progress seems to be very slow.

I hope the influenza will disappear shortly.

Yours sincerely,

W.F. Sheppard.

Letter 5

2 Temple Gardens, E.C.

10 May '99

Dear Mr. Pearson,

The Phil. Mag. would not have the paper, so I have sent it on to the London Math. Society to take its chance.

I have not been doing anything lately as to the quadrature. I had a try at using them for my curves giving correlation-volumes, but found that it was no good, & that it was simpler to calculate the areas by a direct formula. The curves are of the form $z \propto e^{-\mathrm{Csec}^2 \frac{1}{2}x}$, and they have infinitely close contact with the base at the extremities $\pm\pi$.

If you have plenty of spare copies of your paper on correlation of parametric heights, I should be very glad of one.

I suppose you sometimes set examination papers. I hit on a rather nice question the other day (in finding powers of $\tan(67\frac{1}{2}^\circ) = \sqrt{2}+1$): the fundamental formula is not given in Hall & Knight, & only in a generalised form in [Chrystal?]. (I have not got [Serret?]). "Having given the successive powers of $\sqrt{a^n+1}-a$, up to the nth, to a certain number of places of decimals, find a formula for calculating the powers of $\sqrt{a^n+1}+a$ to the same number of places. Hence find the values of $\frac{1}{2}(\sqrt{26}+5)^{10}$ correct to ten places of decimals." or "If $\frac{P_n}{Q_n}$ is the nth convergent to $\sqrt{a^n+1}$, show that $(a+\sqrt{a^n+1})^n = P_n + Q_n\sqrt{a^n+1}$. Hence show that $(5+\sqrt{26})^6$ differs from 1060902 by less than 10^{-6}, and find the value of $\frac{1}{2}(5+\sqrt{26})^{10}$ correct to ten places of decimals." (Ans. 5517851251.00000 00000)

You might make some such question out of it.

I hope you are all right again.

Yours sincerely,

W.F. Sheppard.

Letter 6

To K.P. at Gatwick Surrey

2 Temple Gardens, E.C.

31.3.00

Dear Mr. Pearson,

Have you made any use yet (for any paper to be published) of the quadrature formulae I sent you last year? If not, I should rather like to work them up into a short paper.[5] I had thought of looking up authorities to see what formulae there were in actual use: but that would take time, and it would be simpler to put forward what I have got already. I have looked at a few books, & they never seem to be much beyond Simpson's rule, or the more elementary "trapezoidal rule". I have never come across Parmentier's rule in print: if you can give me a reference to it, & to any others, I should be much obliged.

My long-delayed paper on the calculation of normal-correlation double-integral only went to the Camb. Phil. Soc. a couple of months ago,[6] & I see you have presented a paper to the R.S. on the same subject, but for multiple integral. I wonder whether our methods are at all the same? My paper is now going through the press, but I am not certain when it will be out.

Yours sincerely,

W.F. Sheppard.

Your formula $\frac{d^2z}{dx_{pq}} = \frac{d^2z}{dx_p dx_q}$ is, I suppose, the same as the $x = w.r.\theta$ on p.146 of my Phil. Trans. paper. My formulae are based on this ($\frac{dV}{dD} = -z$).

[5] Sheppard published a paper on quadrature formulae in 1900 titled, "Some quadrature-formulae" in the *Proceedings of the London Mathematical Society* 32, pp. 258–277.

[6] Sheppard is referring to his 1900 paper "On the calculation of the double-integral expressing normal correlation" in *Transactions of the Cambridge Philosophical Society* 19, pp. 23–66.

Letter 7

To K.P. at Gatwick, Surrey

2 Temple Gardens, E.C.

5 April 1900

Dear Mr. Pearson,

Your method certainly seems to be different from mine: I should find it difficult to work out correlation-coeffs with the rapidity you describe.

As regards the quadrature formulae, I do not want you to restrict your use of them: it is only a question of giving them with the proofs. I have just been writing out a sketch of a paper, which, if printed, would run to some 10 or 11 8vo pages, it gives the principal formulae (with extension to calculation of volumes), & the methods by which they are obtained. I could send this to you on your return to town, so that you might compare it with what you have yourself written, & see whether it could be incorporated in your paper.[7]

I hope that in considering the fit of curves to observations you take account of probable error. The method of estimating accuracy of fit by percentages always seems to me unsatisfactory: if 6% is good, for 500 observations, surely it may be bad for 2000? And how do you decide what is good & what is not? What is good for a curve with 3 constants may be bad for a curve with 5 constants. And moreover misfits at different points of the curve are not really of equal weight. Perhaps you have worked out a comprehensive method of dealing with this.

Yours sincerely,

W.F. Sheppard

Letter 8

To K.P. at Gatwick

2 Temple Gardens, E.C.

8 April 1900

Dear Mr. Pearson,

I think that full treatment of the quadrature formulae, with extension to volumes (useful for naval architects etc.), might might[sic] your paper too bulky. However, I should like to discuss the subject with you, and will come to see you any time that

[7]Pearson references Sheppard's quadrature formulae paper 1900 "Some quadrature-formulae" in the *Proceedings of the London Mathematical Society* 32, pp. 258–277 in his 1902 "On the systematic fitting of curves to observations and measurements" in *Biometrika* 1, pp. 265–303.

suits you on April 25th or 26th (or 24th if you prefer it). I will bring what I have on the quadrature-formulae, & on curves of the form *sketch of a concave curve*. (I do not know whether you consider these at all): & also might be able to make some rough notes on measurement of fit. But in this question of fit there are difficulties as to moment-methods, which I should like to discuss with you.

Does your paper touch on interpolation & calculation or ordinates? I have dealt with these, as regards "quasi-normal" curves *sketch of normal curve*, in a recent paper to L.M.S.[8] You may be interested to know that the paper contains very full tables relating to curve of error, abscissa being in terms of S.D.

Remember me to any of the Charterhouse people you may meet: I am sorry to say I have not been down there for two or three years. I hope you find the Shackleford air invigorating. I met an inhabitant the other day, who seemed to have thriven on it: a Miss Walker, I think—an art student in London, but lives with an aunt, I believe, at Shackleford.

I am indoors with a cold, but hope to get away at Easter, to revive.

Yours sincerely,

W.F. Sheppard

Letter 9

2 Temple Gardens, E.C.,

4 May 1900

Dear Mr. Pearson,

My $\theta + \frac{1}{2}$ instead of θ seems quite correct: I enclose proof, written in a slightly different form.

I somehow went off on quite a wrong talk at the end of our discussion today. You are quite right in saying that one does not make an appreciable difference in the fit of a curve of a particular type by altering the constants within the limits they must reasonably have according to the data. My objection is—why take that type? If you have eight classes, and if you take a curve with seven constants, the equation being such that 2 must be positive, you can fit the curve to the data perfectly. But that seems to me to prove nothing. At any rate, it does not prove that the actual curve of frequency is of that type: for if you had taken 16 classes large discrepancies might have appeared. All that you would really be doing, so far as I can tell, would be finding an equation by means of which you could interpolate for intermediate values

[8]Sheppard is referring to his 1900 paper "Some quadrature-formulae" in the *Proceedings of the London Mathematical Society* 32, pp. 258–277.

of the variable: or to put things in my language, you would be finding an anscillary curve which would give a constant first difference. This of course is important, but it does not get you very far on the way to answering the question—what is the law that the frequency in the total "population" follows?

When you take less than $n-1$ arbitrary constants, so that the curve only partially fits the data, it seems to me that there are any number of replies to the question, how to measure the misfit. It is a practical question, depending on the circumstances. If you are dealing with deaths, for the purpose of life assurance, it is much more important to get an accurate fit at the early part of the curve, and any misfits for group lives must be heavily weighted. The relative weights will depend on the rate of interest!

Perhaps I may put my objection more clearly like this. You take, say, 10,000 obsvs. & you get an actual curve *sketch of right-skewed curve with noise* you try the normal curve, & find the misfit puts it out of court. You then try a generalise prob-curve, and find it fits in so closely that *sketch of right-skewed curve with smoothing* the misfit by your method is such as might be due to randomness. In other words, the "wobbles" are actually about a line which does not differ from your theoretical curve by more than the amount permissible by probability-laws. Now take 1,000,000 instead of 10,000. The wobbles, generally speaking will close in to about $\frac{1}{10}$, on the actual line. But how do you know that this closing in will not bring them right away in several places, from your theoretical curve—assuming that the means, mean squares, etc., (up to the order required) are unaltered?

What I should have liked to discuss with you—only it is a very big subject—is the lines on which a more complete paper of the kind you have in hand might go. I think there are three main heads:

I. Manipulation of data without reference to theory as to nature of frequency including smoothing, interpolation, determination of ordinate (true) of the actual frequency-curve, etc. It appears to me that your proposed test of misfit comes under the head of "smoothing". (Possible errors of measurement also come under this.)

II. Investigation of question whether suggested laws of frequency hold; & analysis of the classes of cases for which particular laws hold. My suggested modification of your test comes under this.

III. Dealing with particular data on the assumption of their satisfying certain laws. Here, of course, the bulk of your work as regards variation etc. comes in. But the validity of a good deal of it seems to depend on the previous establishment, under II, of the assumed law. Calcs. of moments etc. are incidental to II and III, not to I. Perhaps this is not very clear.

Yours sincerely,

W.F. Sheppard

Letter 10

2 Temple Gardens, E.C.

7.5.00

"Cloudiness-curve".—Taking range as known to be from 0 to 11, & something slightly ([?] at upper end), I let $M_1 = 7.4487$ (6.9687 for range from $-\frac{1}{2}$ to 10, $M_2 = 76.310$, $\mu_2 = 20.827$. For your Type I this will give $z \propto \frac{1}{x^{.3171}(11-x)^{.9123}}$, which is fairly close to yours. I have not a planimeter, and therefore cannot test closeness of fit.

W.F.S.

Letter 11

2 Temple Gardens, E.C.

18.12.00

Dear Pearson,

I should be very glad to contribute to the "Journal", if I can hit on anything. But just at present I hardly feel sufficiently settled to start anything. I should like some time or other to write an article on the testing of hypothesis: but it involves a good deal of arithmetical work.

I think I could very well get out a short article on interpolation-formulae for surfaces. I should have illustrated my article in the R.S.S. Journal by applying the method to a case of this kind[9] (the example given in my Camb. Phil. Soc. Paper on the $\iint$, in which a double quartile classification has to be deduced from the data[10]): but the article was rather long without it, so I could not put it in. Would this be too technical for "Biometrika"?

Thanks for the coeffs in interpolation-formula. I kept these out of the paper, as I did not want to imperil its being printed.

Yours sincerely,

[9]By the date of this letter, Sheppard had published a paper in the *Journal of the Royal Statistical Society* in 1897 titled, "On the calculation of the average square, cube & c., of a large number of magnitudes" 60, pp. 698–703.

[10]Sheppard is referring to his 1900 paper "On the calculation of the double-integral expressing normal correlation" in *Transactions of the Cambridge Philosophical Society* 19, pp. 23–66.

W.F. Sheppard

Letter 12

163 Kensington Road, S.E.

13.2.01.

Dear Pearson,

I was talking today to Hugh Chisholm, who is editing the Times' supplement to the Encyclopedia Britannica and he told me there was to be no article on mathematical treatment of statistics as he (or his predecessor or superior) had not been able to get your assistance. Don't you think this is a great pity? It would not take you long to write a brief article. Chisholm is very anxious to have you in the list of contributors and the subject ought really to receive mention.

Yours sincerely,

W.F. Sheppard

Edgeworth is writing an article on errors of observation; but this only touches the fringe of it.

Letter 13

30 Oxford Road, Ealing, W.

16.2.02

Dear Pearson,

Thanks for your letter. I am sorry to hear you also are laid up. I should have been inclined to suggest bed rather than sitting over the fire, but no doubt you know what is best!

As to the Tables, I quite understand the impossibilities of producing them in full in Biometrika at present.[11] As to what is the best thing to do, I should be obliged if you would, as you suggest, write to Forsyth. But it is very much a question of who is likely to want to buy them. Originally I rather favoured the idea of the Cambridge Press, with the object of issuing a volume of tables for statisticians (if you felt inclined to support the project) of which this should be a part: but now I rather fight shy of such an undertaking. Also that would have meant 4to, whereas for these tables alone perhaps 8vo would be better.

[11]Sheppard's tables based on the standard normal curve were published in 1903 "New tables of the probability integral" in *Biometrika* 2, pp. 174–190.

But as a single volume by the Camb. Press they might be 8vo. Putting that question aside, I don't very much care how they are produced, & would much rather leave it to your (or Forsyth's) opinion as to the method that would make them most useful. Only I am afraid it would not do to ask for a grant of a definite sum from the B.A.: I can't afford to run the risk of their costing me even 5 pounds to 10 pounds. Assuming then that (when completed) you could help me to get them accepted either by Royal Society or by Camb. Press or by B.A., the question is whether you would like to issue a portion in Biometrika. Do you think Table I would be useful to you if cut down to 5 decimal places? Davenport's little book has a 4-place table $\frac{1}{2}\alpha$: my Phil. Trans. Paper has a 5-place table of z, but by differences of .05 only. A 5-place table seems accurate enough for most practical purposes-'even if no. of observations exceeded 100,000 the inaccuracy is usually a good deal less than the P.E.: & it has the advantage that you need only print the 1st differences (I suppose the derived differences, not the true 1st diffces). I merely make the suggestion as it would enable you to get the table into less space.

As to the other tables, Table II of course gives very little more than Table I for small values of x, but for large values it gives a good deal more information. It is therefore, useful for various purposes—e.g. for calculating moments of the area. (I did think of calculating moments up to the 4th, but doubted if the tables would be worth printing. Edgeworth asked me about such tables, some time ago.) Table III stops at $x = .80$, because of the increasing differences. I have gone throughout on the principle of making tables useful for interpolation (that is the reason, for instance, why I take $\log\frac{1+\alpha}{1-\alpha}$ rather than $\log\frac{1}{1-\alpha}$ as the argument in Table IV): and you will find that Table III if extended much beyond .80 becomes unmanageable, even if the intervals in α are reduced.

Table V you probably don't much care about as you prefer your χ^2 method: but I don't think that question is thoroughly thrashed out yet. I have an idea, by the way, of offering you a short paper later on, on the representation of data by math. formulae, under the two heads (1) best values of constants, assuming a particular form of equation (2) test whether this form may reasonably be accepted as the (or "a") right one. But it wants some thinking about.

There are one or two things I want to ask you about, as I am shamefully behind-hand in my reading. E.g. has any spare mathematician tried to simplify the process of calculating large numerical determinants or is it quite impossible? (I wish Vernon Boys would devise a machine for the purpose.) And have you anywhere gone into the question of what is the "typical" value in non-normal distributions? You like the "mode". But I suppose that sometimes there is a tendency for the correlation-surface in multiple space time i.e. for large no. of correlated observations—to show a simple hump; and, if so, that seems to give the proper typical values for all the observations jointly. Have your data got as far as this?

Yours sincerely,

W.F. Sheppard

Letter 14

Orwell Lodge, Ringstead Road, Sutton, Surrey

13.2.08

Dear Pearson,

Can you tell me what is the etiquette as to using questions in other people's examination-papers for text-book purposes? C.S. Jackson, of Woolwich, is writing a book on elementary dynamics (on modern lines, I suspect); and he says there are a great many questions in University College papers—published annually in the calendar, I suppose—which he would like to utilise. I suppose there is no objection to this? Are these exam-papers published in collected form?

Yours sincerely,

W.F. Sheppard

Letter 15

Braybrooke, Worcester Rd, Sutton, Surrey

18.5.11

Dear Pearson,

It was stupid of me not to see your reasoning: but one soon gets rusty. I do not feel very sure yet.

Generalised, the (incidental) question seems to be this. A, B, C, ... are individuals (people, houses, etc.) m of them, m large. Probability that A is a bachelor is θ_1, probability that B is a widow is θ_2, probability that C is over 6'0" high is θ_3, probability that D is inhabited is θ_4.

These probabilities are supposed not to be independent; is it, e.g. supposed impossible that the replies to the question "Is A a bachelor" etc. should all be "yes". What is the most probable (or mean) no. of affirmative replies to these questions? You, I think, would say $\theta_1 + \theta_2 + \theta_3 + \ldots$; or, at any rate, if $\theta_1 = \theta_2 = \cdots = \theta_4$ then the no. is $m\theta$. I suspect this is right, on the general ground that correlation doesn't come in when we are dealing with 1st moments; though I am rather lazy about it. But will the S.D. from this no. be $\sqrt{m\theta(1-\theta)}$? I feel pretty sure that it won't on account of the correlation.

Now as to the bearings, it is true that, taking any one house by itself, the probability that it has r cases exactly is $\frac{n!}{r!(n-r)!}(\frac{m-1}{m})^{n-r}(\frac{1}{m})^n$; &, this being so for each house, the most probable no. of houses having r is m times above on the argument on preceding page. But the probability for each house is a multiple-integral probability; a kind of average for all the possible combinations of no. in this house with the corresponding no. in the other houses. I don't therefore see how you can treat the case as being the same as if you took the m houses successively with independent probabilities θ in each case. If the 1st house has r, the no. of cases for the other houses is altered from $\frac{(m-1)}{m}r$ to $(n-r)$.

I therefore don't feel that you can treat the question except by taking the possibilities for each case & doing multiple summations, as the authors have done.

Yoursever,

W.F. Sheppard

Letter 16

Sutton

4.10.11

I am not quite sure what tables Everett requires for his further calculations you were mentioning: but are you quite sure none of my M.S. tables would be of use? I have, e.g. to 10 places approx. up to $x = 6.00$ by .01 and tables of x to 7 places, accurate for $\log\frac{(1+\alpha)}{(1-\alpha)}$ up to 6.00.

x	$\log_{10}\frac{1}{2}(1-\alpha)\frac{1}{10}$	Δ
5.98	1.0475428015	266823721
5.99	1.0208604294	267247568
6.00	0.9941356726	267671444
x	$\log_{10}\frac{(1+\alpha)}{(1-\alpha)}$	Δ
5.98	4.7441096	46596
5.99	4.7487691	46554
6.00	4.7534245	46511

W.F.S.

Letter 17

Braybrooke, Worcester Rd, Sutton

23.7.15

Dear Pearson,

I should be very glad to suggest one or two questions, but I feel it rather difficult to estimate the standard required. I will look at the syllabus, and think it over in the next day or two.

Except for my eldest boy, who is away (since March) for open-air cure of nervousness & general weakness, we are all well: & my wife & I are pretty busy. We are moving in Sept. to Berkhamsted (Herts), where all the boys will be able to go to the school as day-boys. I hope that will be our last move, at any rate before I retire, which will not be many years now. Whether my eldest boy will be able to start school at once, I don't know: he may have to have another term or two of the open-air treatment. It certainly has done him good.

What with arragenments for the move & house-getting, some special constable work & occasional drills, I don't have much spare time just now. And I shall get very little when we have moved: the longer journey means an extra hour out of every day.

You don't say anything about yourselves. I hope the boy is getting all right.

Yours sincerely,

W.F. Sheppard

Letter 18

Cardrona, Berkhamsted, Herts

18.10.16

Dear Pearson,

The firm that offered to keep the Brunsvigas in order was Messers J.H. Maxwell & Co., The Albany 21 Mawdsley St. Bolton. Probably these are the same that Elderton referred to.

As regards your problem, I think it is a matter not of further terms in h (the breadth of trapezette) but of further terms depending on the numbers from different parts of the range. Having got *sketch of curve with bins showing n_1, n_2 and so forth* approximate values for the constants, you want to vary them so that the probability of occurrence of $n_1, n_2, n_3, \ldots$ shall be a maximum. I have an idea that this was dealt with in some paper in Biometrika, but I may be wrong. I am afraid I can't solve the problem offhand.

Your sincerely,

W.F. Sheppard

Letter 19

Berkhamsted

10.4.25

Why have a separate antilog table at all? A log table is quite good enough, provided you don't use Everett's formula, e.g. to find N if $\log_{10}$N= .9831868583 etc.: $n = 96203$, $\log_{10} n$ = .983186858386153 etc., $\log_{10}(1-\theta) = -$[5]17569 etc., $\log_e(1-\theta)$ = .[5]40454 etc.: then proceed by successive approx. Of course an antilog table is quicker, but people not accustomed to them are apt to make mistakes, and there is the inference to consider. Hope you will have Easter diner.

W.F.S.

Letter 20

Cardrona, Berkhamsted, Herts

6.9.25

Dear Pearson,

You may remember our discussion of interpolating for logs to several places. I have not seen Thompson's new Table: but I have seen the notice of it in *Nature* for Jan 24, and I still think that the "direct" method of interpolation—i.e. using the derivatives of the function, not the differences—is the best. It not only saves about $\frac{1}{3}$ of the cost of printing (since differences need not be printed at all), but is also quicker.

I have set out on the sheet herewith the process for finding the log of the no. given in Nature, which I suppose is also given in the book itself. You will see that (besides single arithmetic—adding etc.) it involves (1) division of 15 or 18 figs by 5 figs (2) four mult. of 404543598 by 9 figs (3) mult, of 18 figs by 18 figs or, alternatively, mult. of 18 figs by 9 figs & division of the product (which remains on the register) by 9 figs. The whole thing takes about $\frac{1}{4}$ hour on my old Brunsviga: with practice, & a better machine, one might do it quicker.

On the second sheet I mention the continued-fraction convergent method. I have used this a good deal for finding 20-place logs from Callet, who gives only logs of nos. up to 1200. For that purpose it is very useful: so it would be for finding, say, 30-place logs from a table with 5-figure nos. For 20 place logs, 5-figures there is no advantage in it: but it is interesting.

Yours sincerely,

W.F. Sheppard

Letter 21

"Cardrona", Berkhamsted, Herts

26.11.25

Dear Pearson,

Sorry I could not manage to come to the biostats last night. No: I don't want any of my tables returned to me at present. The $\frac{1}{2}\frac{(1-\alpha)}{z}$ is calculated independently.

You don't say whether you want to borrow my existing $\frac{1}{2}\frac{(1-\alpha)}{z}$ table. You are quite at liberty to do so: indeed it would be an advantage, given risk of fire, to have a copy somewhere. The table is 0.0/0.1/10.0 = 24 places approx. (20 certain), but you won't want all that. From the table as it stands I have taken the time of an interpolation to 12 places (i.e. 12 sig. figs.). Working to 12 places, this took 8 mins.: but I was not accustomed to the table. Then I did another, to 12 places but keeping 2 extra figs. in the calculations. This took 7 mins. If you made a copy of the table to 14 places only (12 places plus 2 extra figs), so that there would not be so much hunting for the figures, you ought to be able to do a 12-place interpolation in 5 mins.; which is good enough until a table by intervals of .01 is constructed.

The object of the table was to give a ready means of calculating $\frac{1}{2}(1-\alpha)$. For z one seems to need such tables as $e^{-\theta}$ (1) for $\theta = .000$ to .999 (2) for $\theta = .000000$ to .000 999, etc. I don't know whether these exist. I forget at the moment how Glaisher's & Newman's tables in Trans. Camb. Phil. Soc.[12]

I want to send the table mentioned above to the R.S., even if they don't print it. Action has been hung up in order to investigate degree of accuracy: but I am inclined to drop this, & cut the table down to 20 figs. That is, all that will be wanted in this generation.

I think I told you I have also table of $\log_{10}\frac{1}{2}(1-\alpha)$ to 12 places. I could do it to 16 approx., but it didn't seem worth while. An interpolation would take 6 or 7 mins., as signs have to be studied: then you will have to use a log table. You can borrow this also if you like. Just at the moment I am working on a paper, which might be suitable for Biometrika, on construction of illustrative cases of frequency dists: using a table of 100 values of x taken strictly at random.

[12] J.W.L. Glaisher (1883) "Tables of the Exponential Function" in *Transactions of the Cambridge Philosophical Society* 13, pp. 243–272. F.W. Newman (1883) "Table of the descending exponential function to twelve or fourteen places of decimals" in *Transactions of the Cambridge Philosophical Society* 13, pp. 145–241.

Yours sincerely,

W.F. Sheppard

(missed tonight's post by oversight)

Letter 22

"Cardrona", Berkhamsted, Herts

2.12.25

Dear Pearson

Here are the 3 basic tables I mentioned, with notes on interpolation from them: also 2 others which I do not know whether you have seen.

You can either make such extracts as you like, or keep them by you for the present.

When you return them, you might also return the other tables you have. I think they are 0102, 0115, 0129.

I want to extract the covers & replace them by shorter notes, as they tell me where to look for the rough copies, which I can use when I want them, as I do occasionally now.

I would like to send the 3 tables (0028, 0198, 0199) to the R.S., especially as my Brunsviga is on loan from them & I want them to see the sort of things it is being used for. Would you present them for me, or should I get Whittaker? I could get them ready by some time in Jan.: I am rather tied up now until Xmas.

Thanks for P.C. about Glaisher and Newman. I am inclined to think we need new exponential tables printed: mostly e^x, not e^{-x}. (Eg. To find $\frac{1}{\sqrt{2\pi}}e^{-.123456}$, multiply $\frac{1}{\sqrt{2\pi}}e^{-.124}$ by $e^{+.000544}$: it is much safer.)

Yours sincerely,

W.F. Sheppard

Letter 23

"Cardrona", Berkhamsted, Herts

29.6.26

Dear Pearson,

Thanks: yes, that is the lot. I do hope your op. will be successful.

As regards tables, I don't seem to have any tables of z for the smaller values of x to more than 9 dec. places, & those only approx. What I think you need, in these days of machines, are tables for calculating z by a series of multiplications, not by interpolation. The enclosed sheet shows the scheme. You might put someone on to it!

Yours sincerely,

W.F. Sheppard

Tables for calculating z to (say) 18 figures.

$$\frac{1}{2}x^2 = \theta$$

$$n = \text{integer next above} \quad \theta$$

(1) Preliminary table of $\frac{1}{\sqrt{2\pi}}e^{-n}$ to 12 figures for such (integral) values of n as are wanted: say $n = 0$ to 100.

(2) Table of $e^{\phi} - 1$ to 18 places for $\phi = .000$ to $.999$

(3) Similar table (18 places, = but not greater than 15 figures) for $\phi = .000\ 000$ to $.000\ 999$

(4) Similar table (18 places, = but not greater than 12 figs) for $\phi = .000\ 000\ 000$ to $.000\ 000\ 999$ (for $\phi < .000\ 000\ 000$, $e^{\phi} - 1 = \phi$)

If only 12 figures were catered for, only (1)-(3) would be required, (1)&(2) being to 12 figures & (3) to 9 figures.

Appendix C

Infant Mortality Data

The infant mortality dataset is from the *Seventy-fourth Report* (1911) of the Registrar-General, London, England (1913). The table lists the number of deaths under one year of age for every 1000 live births for 42 years from 1870 to 1911.

Year	*Deaths*	*Year*	*Deaths*	*Year*	*Deaths*
1870	137	1890	135	1910	92
1871	137	1891	136	1911	94
1872	131	1892	133		
1873	131	1893	131		
1874	133	1894	125		
1875	138	1895	133		
1876	128	1896	127		
1877	124	1897	125		
1878	132	1898	123		
1879	127	1899	123		
1880	130	1900	126		
1881	118	1901	119		
1882	128	1902	118		
1883	125	1903	114		
1884	126	1904	115		
1885	127	1905	107		
1886	129	1906	101		
1887	127	1907	105		
1888	125	1908	100		
1889	128	1909	96		

Curriculum Vitae

Name:	Lori Murray
Post-Secondary Education and Degrees:	*A.R.C.T. in Piano and Music Pedagogy, 2000* The Royal Conservatory of Toronto
	Hon. B. Sc. in Mathematical Sciences with Distinction, 2010 The University of Western Ontario
	Master of Science in Statistics, 2012 The University of Western Ontario
Honors and Awards:	Dean's Honor List, 2006-2010 The University of Western Ontario
	Faculty of Science Graduate Teaching Award, 2011 The University of Western Ontario
	Research Poster Award, 2012 Statistical Society of Canada
Teaching:	Introductory Statistics Course, 2012–Present The University of Western Ontario
	R Statistical Software Workshops, February 2016
Papers Presented:	Poster Session, 2012 Award for Best Poster Presentation Statistical Society of Canada

Publications:

Murray, L.L. and Bellhouse, D.R. (2014). A reconstruction of Halley's 1701 map of magnetic declination. *Imago Mundi*, to appear.